Addressing the Challenges of Automated Speech and Language Analysis for the Assessment of Mental Health and Functional Competency

by

Rohit Nihar Uttam Voleti

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved June 2022 by the
Graduate Supervisory Committee:

Visar Berisha, Co-Chair
Julie M. Liss, Co-Chair
Pavan Turaga
Andreas Spanias

ARIZONA STATE UNIVERSITY

August 2022

ABSTRACT

Severe forms of mental illness, such as schizophrenia and bipolar disorder, are debilitating conditions that negatively impact an individual's quality of life. Additionally, they are often difficult and expensive to diagnose and manage, placing a large burden on society. Mental illness is typically diagnosed by the use of clinical interviews and a set of neuropsychiatric batteries; a key component of nearly all of these evaluations is some spoken language task. Clinicians have long used speech and language production as a proxy for neurological health, but most of these assessments are subjective in nature. Meanwhile, technological advancements in speech and natural language processing have grown exponentially over the past decade, increasing the capacity of computer models to assess particular aspects of speech and language. For this reason, many have seen an opportunity to leverage signal processing and machine learning applications to objectively assess clinical speech samples in order to automatically compute objective measures of neurological health.

This document summarizes several contributions to expand upon this body of research. Mainly, there is still a large gap between the theoretical power of computational language models and their actual use in clinical applications. One of the largest concerns is the limited and inconsistent reliability of speech and language features used in models for assessing specific aspects of mental health; numerous methods may exist to measure the same or similar constructs and lead researchers to different conclusions in different studies. To address this, a novel measurement model based on a theoretical framework of speech production is used to motivate feature selection, while also performing a smoothing operation on features across several domains of interest. Then, these composite features are used to perform a much wider range of analyses than is typical of previous studies, looking at everything from diagnosis

to functional competency assessments. Lastly, potential improvements to address practical implementation challenges associated with the use of speech and language technology in a real-world environment are investigated.

The goal of this work is to demonstrate the ability of speech and language technology to aid clinical practitioners toward improvements in quality of life outcomes for their patients.

DEDICATION

First, I dedicate this work to my grandmothers, Rukmani Annapantula and Sarojini Voleti, to whom I owe much more than I could ever put into words. I am so fortunate that both of them were with me when I began this process, and never stopped being my ultimate supporters.

I am even more fortunate that both of my grandfathers, Krishnamurty Annapantula and Rama Rao Voleti, are with me today to continue providing their love, guidance, and support.

Lastly, this is for the rest of my family, without whose support none of this would be possible. Thank you Mom, Dad, Vivek, Divya, and Priya. I love you all.

ACKNOWLEDGMENTS

First and foremost, I would like to acknowledge the mentorship and guidance offered by my co-advisors, Dr. Visar Berisha and Dr. Julie Liss. The opportunity I have had to work with you both is one I never took for granted, and I have learned an immense amount about how to be a better researcher, professional, and person overall. I am fortunate to have felt comfortable enough to also share personal details about struggles or issues independent of research goals, and for that I am forever grateful. Even during moments of personal or academic struggle, I appreciate that you both have always been supportive and available to help me get back on track and succeed. I must also acknowledge the guidance of my other faculty committee members, Dr. Pavan Turaga and Dr. Andreas Spanias. Over the last few years, I appreciate your willingness and enthusiasm to communicate with me about my research, and thank you for always asking insightful questions that helped guide our work. To Dr. Umit Ogras, thank you for being my first mentor at ASU. When our mutual goals were not being met, you were truly supportive and encouraging in helping me find a new place to succeed. At the time, I am not sure I would have had the courage to do so without your guidance.

My time at ASU was made significantly more enjoyable because of my colleagues, who made me a better researcher and provided great friendship and support along the way. From Dr. Ogras' lab, thank you to Ganapati Bhat, Ujjwal Gupta, Cemil Geyik, Md Muztoba, Jaehyun Park, Manoj Babu, Sumit Mandal, Vatika Chaurasia, and Ranadeeb Deb for all of your help and support. From Dr. Berisha and Dr. Liss's lab and Aural Analytics, I would like to offer special thanks to Ming Tu, Yishan Jiao, Megan Willi, Amy LaCross, Vikram C.M., Gabriela Stegmann, Shira Hahn, Pranav Ambadi, Kelvin Tran, Lingfeng Xu, Jianwei Zhang, Weizhi Li, Mazher Khan, Prad

Kadambi, Michael Saxon, Jacob Peplinski, Kan Kawabata, Jennifer Liu, and many others I have had the pleasure to work alongside.

Thank you to the staff at the Fulton Schools of Engineering (ECEE) and the College of Health Solutions at ASU for all of your administrative support over the last several years. Special thanks to my graduate advisors Lynn Pratte and Sno Kleespies for first introducing me to ASU and supporting me throughout with prompt and helpful answers to all my questions. I must also acknowledge the Fulton Schools and ASU Graduate College for their financial support by awarding me the Dean's Fellowship, Fulton Fellowship, and Graduate Completion Fellowship.

Outside of ASU, I would like to specially recognize Dr. Christopher Bowie and his group at Queen's University, without whom much of our research and collaborative publications would have been possible. Within his group, thank you specifically to Stephanie Woolridge and Melissa Milanovic for your contributions to our work, and I hope to continue collaborating more in the near future.

Lastly, I am very fortunate to have wonderful support all throughout my personal life, and there are certainly more individuals who deserve acknowledgment than I will be able to name here. My parents and family have always been able to support me and be a strong safety net throughout this entire process. The same can be said about numerous close personal friends, some of which have been part of my life for more than a decade. To my life partner, Priya, thank you for sticking by me throughout all of this, and I am so happy to have recently married you and look forward to our lives together. Also, thank you so much for bringing Catcat into our lives. Lastly, I am so fortunate to have met my dog and best friend, Dodger, in 2017 while working on my Ph.D., and he has helped take care of me in more ways than I can ever describe. I love you, buddy. Thank you.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

Table Page

Table Page

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

1.1 Background & Motivation

1.1.1 Burden of Mental Illness

Mental illnesses such as major depression, schizophrenia, bipolar disorder, and anxiety disorders are among the most burdensome diseases that have a significant financial and human cost associated with them on a global scale [1]. In the United States alone, the National Institute of Mental Health (NIMH) estimated in 2016 that nearly one in six adults ($\sim$44.6 million people) lives with some form of mental illness [2], and treatment of mental illness may cost approximately $1 trillion annually [3]. By many estimates, the true societal and financial cost is being vastly under-counted and continuing to increase as populations age [1, 4, 5].

Schizophrenia and bipolar disorder (BD) are particularly burdensome, and researchers have identified a critical need to improve our infrastructure for early detection, diagnosis, treatment, and management of these conditions [5, 6]. Analysis of the *Global Burden of Disease* studies from 1990-2017 [7] has shown that BD and schizophrenia impact approximately 4.53 million and 1.13 million people worldwide, respectively, with both conditions showing sharp increases in case incidences as populations continue to age over the past few decades [4, 5]. While prevalence of these conditions may be considered relatively low, the healthcare, social, and financial costs associated with them are disproportionately large and burdensome [6, 8]. For this reason, our work

primarily focuses on the development of computational methods aimed at improving patient outcomes in quality of life and social participation for individuals that are affected by these ailments.

1.1.2 Speech & Language as a Window into Cognitive Health

Many aspects of cognitive and thought disorders are manifest in the way speech is produced and what is said. Irrespective of the underlying disease or condition, the analysis of speech and language can provide insight to the underlying neural function. This has motivated current research trends in quantitative speech and language analytics, with the hope of eventually developing clinically-validated algorithms for better diagnosis, prediction, and characterization of these conditions.

In Figure 1, we see the general procedure researchers typically employ to study cognitive health using speech and language analysis. Patients provide speech samples via a speech elicitation task. This could be passively collected speech, patient interviews, or recorded neuropsychological batteries. The resulting speech is transcribed, using either automatic speech recognition (ASR) or manual transcription, and a set of speech and language features are extracted that aim to measure different aspects of cognitive-linguistic change. As seen in Figure 1, analysis of the extracted language samples can consist of speech signal processing, *i.e.* analysis of the recorded audio signal, or natural language processing (NLP), *i.e.* textual analysis of the linguistic output. These features become the input of a machine learning (ML) model that aims to predict a dependent variable of interest.

Therefore, many researchers have concluded that speech and language output can serve as a useful biomarker for the assessment of cognitive health. A comprehensive

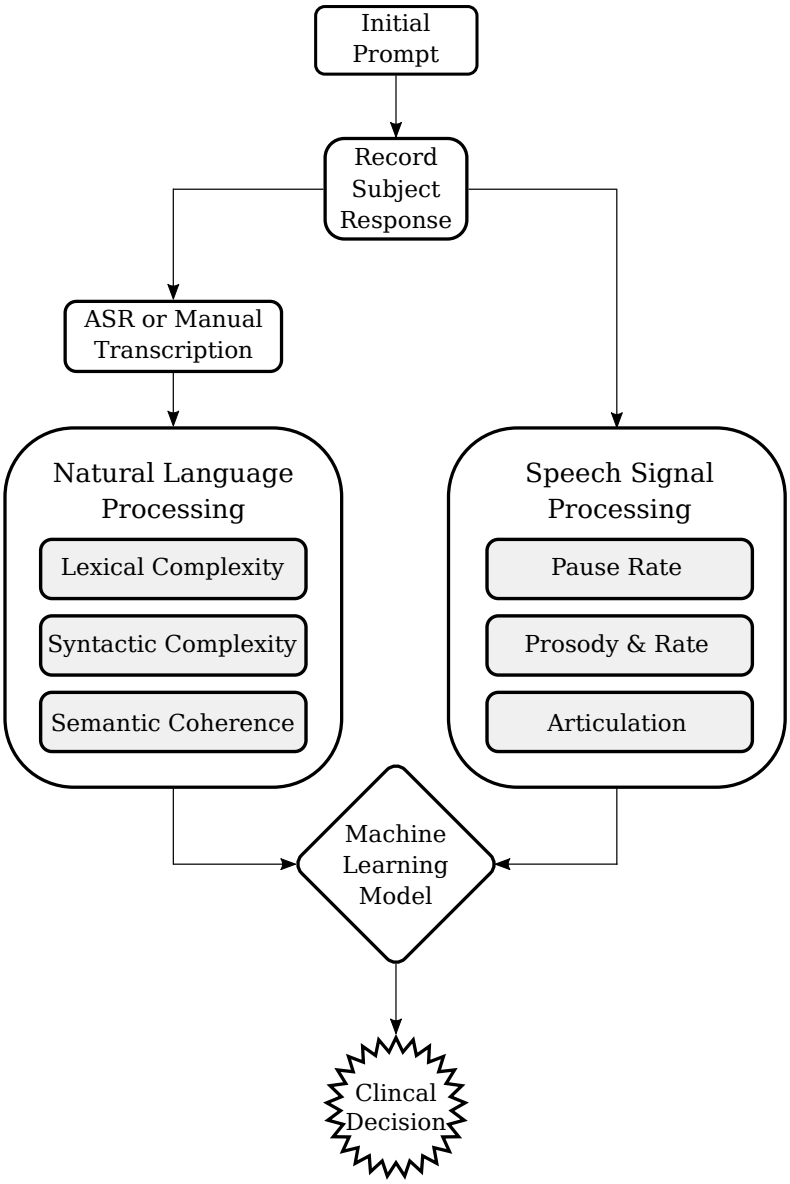


Figure 1: Overview of the process of using natural language processing and speech signal processing for extraction of speech and language features for clinical decision-making. Example language features include lexical complexity, syntactic complexity, semantic coherence, *etc*. Example of acoustic speech features include pause rate, prosody, articulation, *etc*.

review of this work in relation to cognitive and thought disorders can be found in Chapter 2 of this document. However, it is quite clear that there is still a large gap between the latest trends in ML research and actual implementation of computational language analysis in clinical settings [3]. For this reason, our research has particularly focused on improving the practical utility of speech and language processing in clinical settings to improve patient outcomes.

1.2 Sampling of Existing Work

Several papers have established the importance of automated language analytics for assessment of patients with (or at-risk for) thought disorders [9]. As incoherent language is a common symptom across several thought disorders, most of the existing work has focused on computational models of semantically incoherent speech. Here, we provide a brief sampling of the existing literature and highlight our additional contributions, which are formalized in Section 1.3. A much deeper dive which surveys the existing literature using speech processing and NLP to study cognitive and thought disorders is provided in Chapter 2.

One of the earliest studies of language as a predictor of clinical condition [10] primarily focused on formal thought disorder (FTD); the authors compared healthy control participants and those exhibiting FTD by using *latent semantic analysis* (LSA) [11] to generate objective estimates of language similarity scores across samples elicited using a variety of tasks. Bedi *et al.* [12] and Corcoran *et al.* [13] also made use of LSA to predict the onset of psychosis in young individuals deemed to be at clinical high-risk. More recent work has also made use of neural word and sentence embeddings (*i.e. word2vec* [14] and *GloVe* [15]) to assess similar types of

coherence in speech samples from those with schizophrenia or BD [16, 17]. A novel approach using neural word embeddings was recently proposed in [18], in which a *vector unpacking* approach was used to decompose an average sentence vector into its most significant meaning components. The authors show that low *semantic density* for given language elicitation tasks could serve as a reliable predictor for the onset of psychosis. Beyond semantics, other aspects of language have been computationally analyzed for individuals with schizophrenia and BD. For example, previous work has measured different features related to syntax [19], conversational pragmatics [20], several measures of language complexity [17], ambiguous pronouns [16], among others.

1.3 Problem Statement, Research Summary, and Contributions

Most of this previous work has taken a data-driven approach to identifying language metrics as useful prognostic and diagnostic markers for schizophrenia and BD. However, this technological potential to improve patient outcomes is currently unrealized in clinical practice [3]. While this literature clearly shows that there is value in language analytics for assessing thought disorders, the studies are fragmented with little to no standardization. For example, nearly every study uses a completely different set of metrics for measuring the same (or similar) constructs. Similarly, speech and language samples from clinical populations are often specific to a particular task and are not widely available in a standard format, further complicating this issue. Developers therefore tend to take a data-driven approach in developing machine learning models that are optimized for regression or classification performance based on what data they have available, but these models are not always clearly interpretable.

Another problem is that existing literature in the field mostly focuses on evaluating

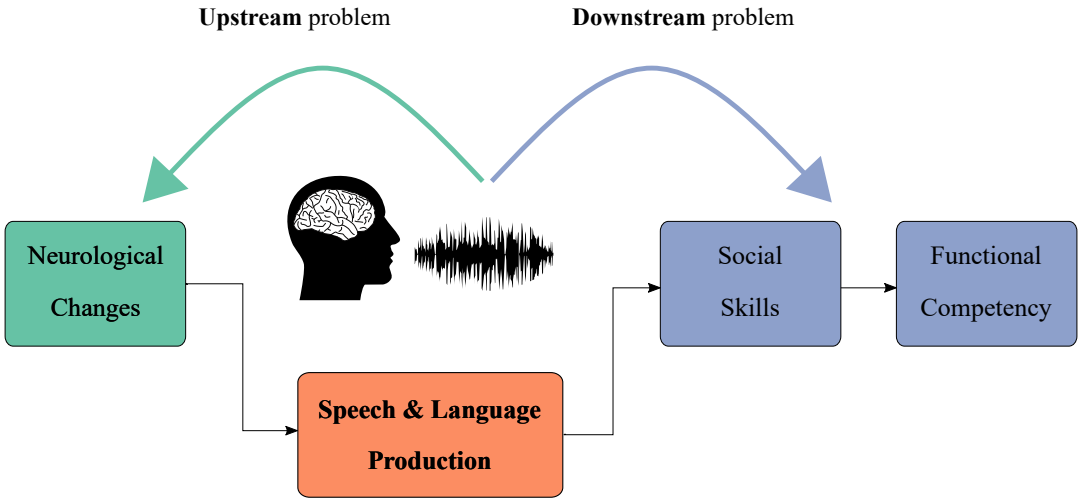


Figure 2: Speech and language characteristics are manifestations of the underlying upstream neurological changes the patient is experiencing. Speech and language abnormalities have important downstream consequences on activities of daily living and participation (*e.g.* social interactions, work, *etc.*). In this document, we define the upstream problem as using speech and language analysis for diagnosis or prognosis, and we define the downstream problem as using speech and language analysis for assessment of social and functional competency.

the ability of speech and language analysis to detect or identify the presence of an underlying neurological condition or mental illness; we call this the *upstream* problem of clinical diagnosis or symptom assessment. While this is certainly valuable information to practitioners, we argue it is at least equally important for practitioners to assess outcomes in social and functional competency, especially when evaluating the impact of interventions; we call this the *downstream* problem. Recorded speech output is our starting point to address both of these issues in our schizophrenia and BD data set. This concept is summarized in Figure 2.

Therefore, our research expands upon previous work to use speech and language to address both of these issues within a clinically interpretable theoretical frame. We address the fragmentation and interpretability problem by motivating our selection of

speech and language metrics from a theoretical framework for speech and language production, first introduced in Chapter 2. We reduce all of our computed features to a lower-dimensional representation that focuses on clinical domains of interest. Additionally, we test the ability of this interpretable language feature set to address both the upstream and downstream problems outlined in Figure 2.

In summary, our research aims to accomplish the following to improve the adoption of computational language analysis in clinical practice:

1. Development of a *measurement model* for speech and language analysis based on a theoretical framework of how speech and language are produced; the purpose of this model is to allow
2. Using the framework from aim (1), we develop interpretable language measures to improve early detection, assessment, and diagnosis of these afflictions (the *upstream* problem), while also using the same measures to assess real-world outcomes in social and functional competency (the *downstream* problem)
3. Address challenges for real-world implementation of computational language analysis in clinical settings, particularly with regard to noisy data due to automatic speech recognition (ASR).
4. Explore the use of state-of-the-art language generation models to create viable synthetic data with which we can train more robust clinical models.

Our work is conducted in collaboration with Dr. Christopher Bowie and his group in the Psychology department at Queen's University in Kingston, ON, Canada. Their group has provides us with an extensive dataset of deeply-phenotyped patients diagnosed with schizophrenia/schizoaffective disorder and bipolar disorder, as well as a smaller cohort of healthy controls. Among the available neuropsychological tests are the transcripts for a series of role-playing interviews to evaluate social skills for

each participant, part of what is known as the *Social Skills Performance Assessment* (SSPA) [21]. In this work, we have used this data set to evaluate computational natural language processing models to do a series of evaluations to address the upstream and downstream problems described above using these SSPA transcripts.

1.3.1 Contributions to Date:

The measurement model using the theoretical Levelt framework for speech and language production was first introduced in the context of cognitive and thought disorders in our review article published in the IEEE Journal of Selected Topics in Signal Processing in 2020 [9]. A preliminary study was published prior to this [17], in which we used a set of SSPA transcripts to identify language features that serve as good predictors for predicting their evaluation scores. Because these measures were validated by a well-known functional assessment, we argue that they contain clinically relevant information which serves as a basis for correctly identifying participant types by building ML models based on these metrics. Our later work expands upon on this preliminary study with a much larger sample of participants [22]. To further support aim 1, we motivate our feature selection based on the Levelt framework and measurement model for speech and language feature assessment that we proposed in [9]. We also conduct a more robust analysis using these language measures, in which we address both the "upstream" and "downstream" problems mentioned in aim (2). The impact of neurocognitive deficits on real-world functional outcomes is known to be mediated by their impact on social skills and measurable symptoms [23]. We use computational analysis of the language in the transcripts to quantify the social skill

performance and study its ability to predict linguistic features of schizophrenia and BD and its impact on other real-world functional assessments.

In aim (3), we attempt to improve the performance of computational language modeling for use in clinical applications by addressing real-world implementation challenges. Manual transcription of spoken language is a cumbersome task that limits the use of computational linguistic analysis, and any wide-scale adoption would likely be dependent upon automatic speech recognition (ASR). However, ASR is error-prone, and it is currently unclear how transcription errors may impact our downstream analysis. For this reason, we proposed a method to simulate realistically plausible ASR errors in [24], and we evaluated the performance of popular language modeling techniques for semantic similarity when errors were introduced. Future work involves improving language modeling to be robust to the types of errors introduced by ASR to make it more viable for real-world applications.

The data collection process also requires a significant human effort prior to recording and transcription. For this reason, in aim (4), we document the development and use of language generating models at the forefront of artificial intelligence to generate synthetic data to help improve model robustness. We also outline the development and results of conversations between participants and an AI-powered chatbot that was trained to play the role of a clinical evaluator for the SSPA tasks. As this chatbot has conversations with more individuals, its contextual awareness will improve. Additionally, the language samples we collect from participants can be used to further refine and validate our language metrics and the models we develop in aims (1)-(2).

1.4 Potential Impact

The work described herein has the potential for significant impact. It is understood that the overall goal of medical or psychological interventions and treatments is to improve quality of life for individuals afflicted with these ailments. Computational speech and language analysis can aid the development of interventions that improve patient outcomes in both the near and distant future. In the nearer-term, there is the potential for improving the efficiency of clinical trials evaluating new drugs. It is generally accepted that early enrollment in clinical trials for evaluation of new drugs maximizes the chances of showing that a drug is successful [25, 26]. In addition, adopting endpoints that are more sensitive to change means that these studies can be powered with fewer participants. Digital endpoints collected frequently have recently garnered interest in this domain [27]. In the long-term, there is the potential for new diagnostics and early interventions for improved treatment outcomes. For example, a common issue reported by clinical providers is the prevalence of suicide and other severe negative outcomes for individuals who stop taking essential medications or seeing their providers to seek help. If we are able to overcome implementation challenges to conduct real-time monitoring of speech through wearable or mobile devices, speech and language provide a uniquely effective and minimally invasive modality for observing these individuals outside of the clinic to trigger necessary interventions. The ability to effectively evaluate downstream impacts of mental illness in social and functional competency tasks is critical for this type of intervention; therefore, studying the potential of computational speech and language analysis tools to reliably make these assessments is an essential and worthy endeavor.

1.5 Organization of This Document

The rest of this dissertation is organized as follows:

- Chapter 2 provides an overview of the existing literature in the field of using speech processing and NLP to evaluate cognitive and thought disorders. Additionally, we present some of our preliminary work in this area with a sample of the schizophrenia and BD data set used in the rest of our work.
- Chapter 3 provides a motivation for our selection of linguistic features using a theoretical model for speech and language production. We also provide implementation details for the methods used to compute linguistic features from the transcripts in our study, and include the procedures for developing the prediction models used in our work.
- Chapter 4 provides a deeper analysis of the *upstream* and *downstream* problems described above. We use computational language metrics to develop models to predict measures of overall mental health status and social and functional competency. Upstream models include assessment of neurocognition, symptom ratings, and diagnostic classification experiments, while we use language metrics to predict clinical measures of social and functional competency for the downstream problem, *i.e.* proxies for measuring quality of life and social participation.
- Chapter 5 delves into some of the real-implementation challenges that exist in adopting automated computational speech and language analysis in clinical settings. We give a summary of some initial work in this area about dealing with ASR substitution errors in language modeling, and some inital analysis of

what we would like to do in order to make computational speech and language analysis more viable in the mental health clinic.

- Chapter 6 covers a series of experiments we conducted in pursuit of aim (4), in which we generate synthetic conversational data to aid in the improvement of our prediction models with the use of language generating deep neural networks.
- Chapter 7 concludes the dissertation by summarizing our work, describing its place in the greater context of this field, and proposing suggestions for future studies.

Chapter 2

EXISTING RESEARCH AND PRELIMINARY WORK

In this chapter, we provide an overview of the common methods and previous work that use speech signal processing and natural language processing in order to assess a wide range of cognitive and thought disorders through the analysis of speech or verbal output.

2.1 Background

2.1.1 Spoken Language Production

The production of spoken language in humans is a complex, multi-stage process that involves high levels of memory, cognition, and sensorimotor function. There are three distinct stages [29]:

1. *Conceptualization*: involves the formation of abstract ideas about the intended message to be communicated
2. *Formulation*: involves forming the exact linguistic construction of the utterance to be spoken
3. *Articulation*: involves actually producing sounds using the various components of the speech production system, *i.e.* lungs, glottis, larynx, vocal tract, *etc.*

These stages are visually represented in the block diagram in Figure 3. In the conceptualization stage, pre-verbal ideas are formed to link a desired concept to be expressed to the spoken language that is eventually formed. The formulation

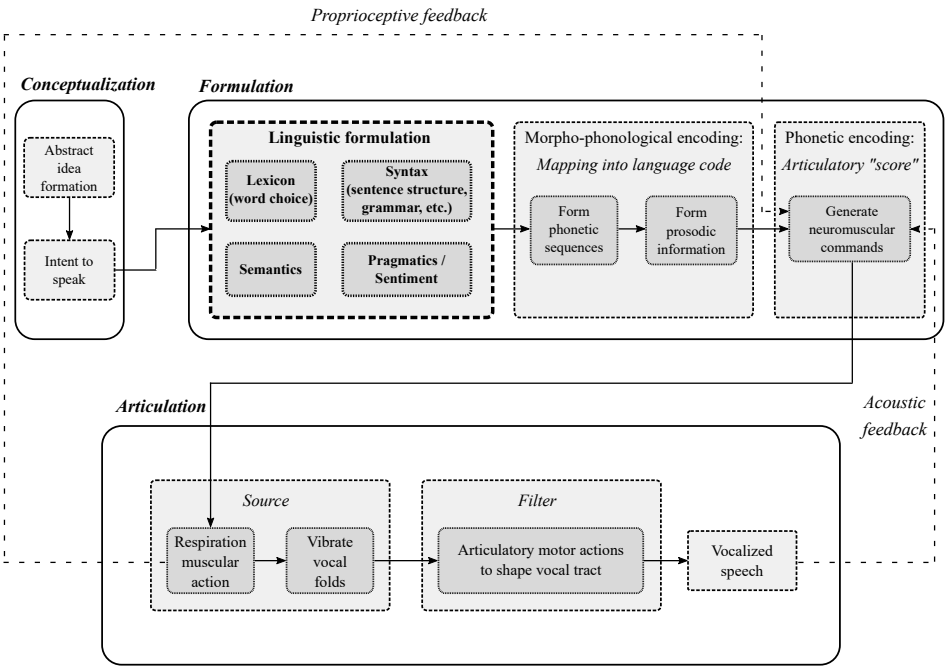


Figure 3: Speech production block diagram model, adapted and modified from [28]. In this review, we focus primarily on the additional box we termed "Linguistic Formulation" within the formulation stage of speech production. Cognitive and thought disorders that affect this area have direct measurable outputs on the actual language content which can be studied by statistical text-based analysis. Additionally, they also have indirect downstream effects on the vocalized and articulated speech acoustics. Both of these areas are covered in our review.

stage consists of several distinct components: (a) lexical, syntactical, & grammatical formulations, (b) morpho-phonological encoding, and (c) phonetic encoding. This involves forming the linguistic structure of a spoken utterance, determining which syllables are needed to articulate the utterance, and the creation of an *articulatory score* containing instructions that are to be executed by the vocal apparatus in the articulation stage [29].

Cognitive and thought disorders have the ability to affect any of these stages, but broadly, they can be captured through analysis of "content" (what is said) and

"form" (how it is said). Indeed, the tools used to characterize content and form of speech are agnostic to the underlying condition. It is the constellation of features shown to be affected that converge on the locus of deficit for an individual. For example, speech that lacks coherence of ideas and jumps from topic to topic (impaired content), and is produced very rapidly and without pauses (impaired form), would point toward a thought or mood disorder, such as schizophrenia or mania. A person with dementia may present with reduced vocabulary size (impaired content), and with increased number and duration of pauses (impaired form). To reiterate, the speech and language measures are, themselves, agnostic to the underlying disorder. Rather, it is the constellation of deficit patterns that associate with different etiologies. Our aim here is to provide an overview of the methods used to extract these constellations without focusing on a particular disease area. We refer to clinical applications in each section to highlight existing work that uses these features in clinical applications.

2.1.2 Clinical Assessment of Speech & Language for Cognitive & Thought Disorders

A variety of clinical protocols exist for the evaluation and diagnosis of disorders affecting cognitive function in psychiatry and neurology. The DSM-5 [30], published by the American Psychiatric Association (APA), provides the standard diagnostic criteria for psychiatric and neurocognitive disorders, and it is updated as knowledge in the field evolves. The DSM-5 covers a large spectrum of psychiatric and cognitive disorders, such as depression, anxiety, schizophrenia, bipolar disorder, dementia, and several more.

Based on these criteria, many evaluation methodologies have been developed in clinical practice for diagnosing and evaluating these cognitive disorders. The *mental*

status examination (MSE) is a commonly utilized and multi-faceted tool for screening an individual at a given point in time for signs of neurological and psychiatric disorders [31]. Components of the MSE evaluate affect, mood, appearance, judgment, speech, thought process, and overall cognitive ability through a variety of tasks and surveys. Related screenings include the *mini-mental state examination* (MMSE) [32], *Addenbrooke's Cognitive Examination* (ACE) [33], and *the Montreal Cognitive Assessment* (MoCA) [34] for evaluating conditions like mild cognitive impairment (MCI), dementia, and Alzheimer's disease (AD). Other forms of disorders, *i.e.* schizophrenia and bipolar disorder, can be evaluated with screenings such as the *Clinical Assessment Interview for Negative Symptoms* (CAINS) [35], *Brief Negative Symptom Scale* (BNSS) [36], the *Social Skills Performance Assessment* (SSPA) [21], and several others that measure the effects of thought and mood disorders.

All of these neuropsychological batteries for evaluating cognitive health have a significant speech and language assessment, as cognitive-linguistic function is a strong biomarker for neuropsychological health in many dimensions. However, ratings for narrative, recall, conversational, or other spoken language tasks are often subjective in nature and of variable reliability, making the underlying diagnosis more challenging [37]. While the diagnosis for many common psychiatric conditions has become more consistent over time as they are better understood, others (*e.g.* schizoaffective disorder) are often evaluated inconsistently by different clinical assessors due to the subjective nature of the test batteries applied [38, 39]. The speech and language samples collected during these screenings serve as potentially valuable databases for *objective* and automatically computable measures of cognitive-linguistic ability. Recent research suggests that analysis of this rich data allows us to explore several new objective dimensions for evaluation, which has a largely untapped potential to improve clinical

assessment and outcomes. These new tools have the potential to provide a finer-grained analysis of the resultant speech when compared against existing rating scales.

2.1.3 Speech & Language Dimensions of Interest

Natural spoken language contains several measurable dimensions that indicate various aspects of cognitive health. In this review, we are interested in the analysis of linguistic and acoustic speech features that are indicative of cognitive and thought disorders related to *thought content* and *thought organization*. These include a variety of neurological impairments (*e.g.* MCI, dementia, AD, chronic traumatic encephalopathy) and psychiatric conditions (*e.g.* schizophrenia, schizoaffective disorder, bipolar disorder).

Most of the work in this space exists in the context of textual language analysis, either by manual or automatic transcription of spoken utterances. Looking at Figure 3, we focus mainly on the "*Linguistic formulation*" area within the formulation stage. Neurological thought disorders all affect the ability of an individual to form complex thoughts and sentence structures, and may often have issues such as *poverty of speech* or disorganized speech. Therefore, we look at methods for examining thought content density, complexity of sentence syntax, semantic coherence, and sentiment analysis as they relate to these conditions.

Analysis of acoustic speech samples leads to additional insight for characterizing neurological and psychiatric thought disorders, as impairments in language formation in turn affect the articulation of the spoken output. As seen in Figure 3, the articulation pathway that leads to speech output depends upon the cognitive ability required for the conceptualization and formulation stages of speech production. Cognitive and working

memory disorders can lead to impairments in neuromuscular motor coordination and proprioceptive feedback as well, affecting speech output [40]. Among the speech signal features considered are those related to temporal analysis and prosody (*e.g.* pause rate, phonation rate, periodicity of speech, *etc.*) and those related to frequency analysis (*e.g.* mean, variance, kurtosis of Mel frequency cepstral coefficients).

We note that the purpose of this review is to highlight recent research that identifies and characterizes automatically computed speech and language features related to neurological and psychiatric disorders of thought content and formulation. In each part, we will provide an overview of commonly used techniques for extracting various speech and language features, present examples of their clinical application, and discuss the advantages and disadvantages of the methods reviewed.

2.2 Measuring Cognitive and Thought Disorders with Natural Language Processing

In this section, we will provide a review of several families of natural language processing methods that range from simple lexical analysis to state-of-the-art language models that can be utilized for clinical assessment.

The sections below present families of approaches in order of increasing complexity. In the first section, we describe methods based on subjective evaluation of speech and language; then we discuss methods that rely on lexeme-level information, followed by methods that rely on sentence-level information, and end with methods that rely on semantics. For each section, we provide a description of representative approaches and a review of how these methods are used in clinical applications. We end each section with a discussion of the advantages and disadvantages of the approaches in that section.

2.2.1 Early Work

Simple analysis of written language samples has long been thought to provide valuable information regarding cognitive health. One of the best-known early examples of such work is the famous "nun study" by Snowdon *et al.* on linguistic ability as a predictor of Alzheimer's disease (AD) [41]. In this work, manual evaluations of the linguistic abilities of 93 nuns were conducted by analysis of autobiographical essays they had written earlier in their lives. The researchers evaluated the linguistic structure of the essays by scoring the grammatical complexity and idea density in the writing samples. In particular, the study found that low idea density in early life was a particularly strong predictor of reduced cognitive ability or the presence of AD in later life. Roughly 80% of the participants that were determined to lack linguistic complexity in their writings developed AD or had mental and cognitive disabilities in their older age.

This work was groundbreaking in showing that linguistic structure and complexity can serve as a strong predictor for the onset of AD and potentially other forms of cognitive impairment. However, it required tedious manual analysis of writing samples and careful consideration that the scores given by different evaluators had a high correlation, due to the subjective nature of the scoring.

These factors make in-clinic use prohibitive; as a result, these methods have received limited attention in follow-on work. The development of *automated* and *quantitative* metrics to analyze language complexity can potentially save several hours of research time to conduct similar linguistic studies to understand neurodegenerative disease and mental illness. Several techniques devised in the NLP literature have been utilized

to address the challenge of conducting quantitative analysis to replace traditionally subjective and task-dependent methods of measuring linguistic complexity.

2.2.2 First Order Lexeme-Level Analysis

2.2.2.1 Methods

Automated first-order lexical analysis, *i.e.* at the lexeme-level or word-level, can generate objective language metrics to provide valuable insight into cognitive function. The most basic approaches treat a body of text as a *bag of words*, meaning the ordering of words within the text is not considered. This can be done by simply considering the frequency of usage of particular words and how they relate to a group of individuals. Specialized tools, such as *Linguistic Inquiry and Word Count* (LIWC) [42], are often used to analyze the content and categorize the vocabulary within a text. LIWC associates words in a text with categories associated with affective processes (*i.e.* positive/negative emotions, anxiety, sadness, *etc.*), cognitive processes (*i.e.* insight, certainty, causation, *etc.*), social processes (*i.e.* friends, family, humans), the presence of dysfluencies (pauses, filler words, *etc.*), and many others. The categorization of the lexicon allows for further tasks of interest, such as sentiment analysis based on the emotional categories. The frequency of usage and other statistics of words from particular categories can lend insight to overall language production.

The concept of *lexical diversity* refers to a measure of unique vocabulary usage. The *type-to-token ratio* (TTR), given in Equation (2.1), is a well-known measure of lexical diversity, in which the number of unique words (*types*, V) are compared against

the total number of words (*tokens*, N).

$$\text{TTR} = \frac{V}{N} \tag{2.1}$$

However, TTR is negatively impacted for longer utterances, as the diversity of unique words typically plateaus as the number of total words increase. The moving average type-to-token ratio (MATTR) [43] is one method which aims to reduce the dependence on text length by considering TTR over a sliding window of the text. This approach does not have a length-based bias, but is considerably more variable as the parameters are estimated on smaller speech samples. *Brunét's Index* (BI) [44], defined in Equation (2.2), is another measure of lexical diversity that has a weaker dependence on text length, with a smaller value indicating a greater degree of lexical diversity,

$$\text{BI} = N^{V^{-0.165}}. \tag{2.2}$$

An alternative is also provided by *Honoré's Statistic* (HS) [45], defined in Equation (2.3), which emphasizes the use of words that are spoken only once (denoted by V_1),

$$\text{HS} = 100 \log \frac{N}{1 - {}^{V_1}\!/_{V}}. \tag{2.3}$$

The exponential and logarithm in the BI and the HS reduce the dependence on the text length, while still using all samples to estimate the diversity measure, unlike the MATTR.

Measures of *lexical density*, which quantify the degree of information packaging within an utterance, may also be useful for cognitive assessment. *Content words*[1] (*i.e.* nouns, verbs, adjectives, adverbs) tend to carry more information than *function*

[1]Content words are also referred to as "open-class", meaning new words are often added and removed to this category of words as language changes over time.

words[2] (*e.g.* prepositions, conjunctions, interjections, *etc.*). These can be used to compute notions of *content density* (CD) in written or spoken language, given in Equation (2.4),

$$\text{CD} = \frac{\#\text{ of verbs} + \text{nouns} + \text{adjectives} + \text{adverbs}}{N}. \tag{2.4}$$

Part-of-speech (POS) tagging of text samples is one way in which the word categories can be automatically determined; individual word tokens within a sentence are identified and labeled as the part-of-speech that they represent, typically from the Penn Treebank tagset [46]. Several automatic algorithms and available implementations exist for rule-based and statistical taggers, *i.e.* using a *hidden Markov model* (HMM) or *maximum entropy Markov model* (MEMM) implementation to determine POS tags with a statistical sequence model [47]. For example, the widely-used *Stanford Tagger* [48] uses a bidirectional MEMM model to assign POS tags to samples of text. Several notions of content density can be computed at the lexeme-level if POS tags can be automatically determined to reflect the role of each word in an utterance. Examples of these include: the *propositional density* (P-density), a measure of the number of expressed propositions (verbs, adjectives, adverbs, prepositions, and conjunctions) divided by the total number of words, and the *content density*, which is a measure of the ratio of content words to function words [49, 50].

2.2.2.2 Clinical Applications

Several studies have utilized first order lexical features to assess cognitive health by automated linguistic analysis. The simplest bag-of-words analysis for vocabulary

[2]Function words are also referred to as "closed-class" since words are rarely added to or removed from these categories.

usage can often provide valuable insight in this regard. For example, the work by Garrard *et al.* computed vocabulary statistics for participants with left- ($n = 21$) and right-predominant ($n = 11$) varieties of semantic dementia (SD) and, and compared them with language samples from healthy controls ($n = 10$) [51]. Classification accuracy of over 90% was reached for categorizing the participants for two tasks: (1) participants with SD against the healthy control participants, and (2) classifying the left- and right-predominant variants of SD. They used the concept of information gain to determine which word types were most useful in each classification problem. Asgari *et al.* used the LIWC tool [42] to study the language of those with *mild cognitive impairment* (MCI), often a precursor to Alzheimer's disease (AD) [52]. The transcripts of unstructured conversation with the study's participants were analyzed with LIWC to generate a 68-dimensional vector of word counts that fall within each of the 68 subcategories in the LIWC lexicon. They were able to achieve over 80% classification accuracy by selecting LIWC categories that best represented the difference in the MCI and healthy control datasets.

Roark *et al.* considered a larger variety of speech and language features to detect MCI [49]. In this work, the authors compared the language of elderly healthy control participants and patients with MCI on the Wechsler Logical Memory I/II Test [53], in which participants are tested on their ability to retell a short narrative that has been told them at different time points[3]. Among the features considered included multiple measures of lexical density. POS tagging was performed on the transcripts of clinical interviews of patients with MCI and healthy control participants. Two measures of lexical density derived from the POS tags were the P-density and the content density. In particular, the content density was a strong indicator of group differences between

[3]Asked to retell the story immediately (LM1) and after approximately 30 minutes (LM2)

healthy controls and patients with MCI. The automated language features were used in conjunction with speech features and clinical test scores to train a support vector machine (SVM) classifier that achieved good leave-pair-out cross validation results in classifying the two groups (AUC $= 0.732,\ 0.703,\ 0.861$ when trained on language features, language features + speech features, and language + speech features + test scores, respectively)[4].

Bucks *et al.* [54] and Fraser *et al.* [50] both used several first-order lexical features in their analysis of patients with AD. In [54], the authors successfully discriminated between a small sample of healthy older control participants ($n = 16$) and patients with AD ($n = 8$) using TTR (Equation (2.1)), BI (Equation (2.2)), and HS (Equation (2.3)) as measures of lexical diversity or vocabulary richness. They additionally considered the usage rates of other parts of speech (*i.e.* nouns, pronouns, adjectives, verbs). In particular, TTR, BI, verb-rate, and adjective-rate all indicated strong group differences between the participants with AD and healthy controls; the groups could be classified with a cross-validation accuracy of 87.5%. Fraser *et al.* [50] performed similar work using the *DementiaBank*[5] database to obtain patient transcripts. They additionally used other vocabulary-related features, such as frequency, familiarity, and imageability values for words in the transcripts. This work was in turn based on a previous study [55] in which similar features were extracted to study the language of participants with two different subtypes[6] of primary progressive aphasia (PPA) and healthy control subjects.

[4]Additional language and speech features will be discussed later

[5]https://dementia.talkbank.org/access/, Accessed August 20, 2019

[6]Progressive nonfluent aphasia (PNFA) and semantic dementia (SD)

Berisha *et al.* performed a longitudinal analysis of non-scripted press conference transcripts from U.S. Presidents Ronald Reagan (who was diagnosed with AD late in life) and George H.W. Bush (no such diagnosis) [56]. Among the linguistic features that were tracked were the lexical diversity and lexical density for both presidents over several years worth of press conference transcripts. The study shows that the number of unique words used by Reagan over the period of his presidency steadily declined over time, while no such changes were seen for Bush. These declines predated his diagnosis of AD in 1994 by 6 years, suggesting that these computed lexical features may be useful in predicting the onset of AD pre-clinically. A related study examined the language in interview transcripts of professional American football players in the National Football League (NFL) [57], at high-risk for neurological damage in the form of chronic traumatic encephalopathy (CTE). The study longitudinally measured TTR (Equation 2.1) and CD[7] (Equation 2.4) in interview transcripts of NFL players ($n = 10$) and NFL coaches/front office executives[8] ($n = 18$). Previous work has shown that TTR and CD are expected to increase or remain constant as healthy individuals age [58, 59, 60]. However, this study demonstrated clear longitudinal declines in both variables for the NFL players while showing the expected increase in both variables for coaches and executives in similar contexts.

[7]The authors in [57] refer to CD simply as "lexical density" (LD)

[8]Coaches and executives were limited to those who were not former players experiencing similar head trauma to serve as a control in the language study.

2.2.2.3 Advantages & Disadvantages

It is clear from the literature that first-order lexeme-level features, *i.e.* those related to lexical diversity and density, are useful biomarkers for detecting the presence or predicting the onset of a variety of conditions, such as MCI, AD, CTE, and potentially several others. POS tagging has several reliable and accurate implementations, and these features are simple and easy to compute. Additionally, these linguistic measures are easily clinically interpretable for measuring cognitive-linguistic ability.

However, lexeme-level features are limited in what information they provide alone, and many of the previously discussed works used these features in conjunction with several other speech and language features to build their models for classification and prediction of disease onset. Since these measures are based on counting particular word types and tokens, they tell us little about how individual lexical units interact with each other in a full sentence or phrase. Additionally, measures of lexical diversity and lexical density provide little insight regarding semantic similarity between words within a sentence. For example, the words "car", "vehicle", and "automobile" are all counted as unique words, despite there being a clear semantic similarity between them[9] In the following sections, we will discuss more complex language measures that aim to address these issues.

[9]Note: lexical diversity is still a potentially useful measure in this case, as a diverse word choice may indicate higher cognitive function.

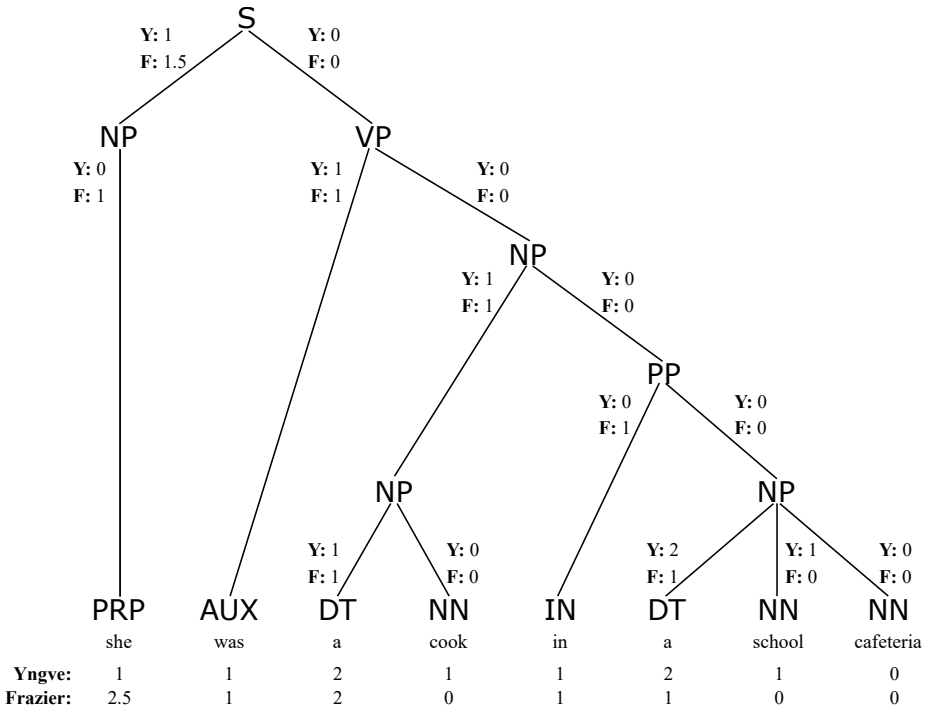
2.2.3 Sentence-Level Syntactical Analysis

Generating free-flowing speech requires that we not only determine which words best convey an idea, but also to determine the order in which to sequence the words in forming sentences. The complexity of the sentences we structure provides a great deal of insight into cognitive-linguistic health. In this section we provide an overview of various methods used to measure syntactic complexity as a proxy for cognitive health.

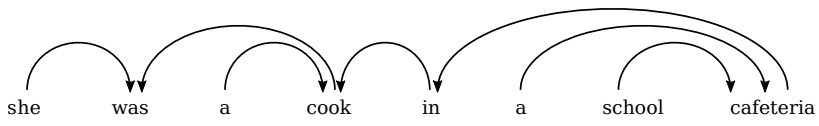
2.2.3.1 Methods

The ordering of words in sentences and sentences in paragraphs can also provide important insight into cognitive function. Many easy-to-compute and common structural metrics of language include the mean length of a clause, mean length of sentence, ratio of number of clauses to number of sentences, and other related statistics [50]. Additionally, several more complicated methods for syntactical analysis of natural language can also be used to gain better insight for assessing linguistic complexity and cognitive health.

A commonly used technique involves the *parsing* of naturally produced language based on language-dependent syntactical and grammatical rules. A *constituency-based parse tree* is generated to decompose a sentence or phrase into lexical units or tokens. In English, for example, sentences are read left to right and are often parsed this way. An example of a common constituency-based left to right parse tree can be seen in Figure 4a for the sentence "She was a cook in a school cafeteria", adapted from [49]. At the root node, the sentence is split into a *noun phrase* ("she") and a *verb phrase* ("was a cook in a school cafeteria"). Then, the phrases are further parsed



(a) Constituency-based parsing of sample sentence (*i.e.* top-down and left to right). In the diagram, S = sentence, NP = noun phrase, VP = verb phrase, PP = prepositional phrase, PRP = personal pronoun, AUX = auxiliary verb, DT = determiner, NN = noun, and IN = preposition. The figure contains examples of both Yngve scoring (Y) [61], Frazier scoring (F) [62] for each branch of the tree. At the bottom is the total score of each type for each word token in the sentence summed up to the root of the tree.



(b) Dependency-based parsing of the same sample sentence. Lexical dependency distances can be computed.In this example, there are 7 total links, a total lexical dependency distance of 11, and an average distance of $^{11}/_{7} = 1\frac{4}{7}$. Longer distances indicate greater linguistic complexity.

Figure 4: (a) A *constituency-based* and (b) *dependency-based* parsing of a simple sentence. Both adapted from [49].

into individual tokens with a grammatical assignment (nouns, verbs, determiners, *etc.*). Simple sentences in the English language are often *right-branching* when using constituency-based parse trees. This means that the subject typically appears first and is followed by the verb, object, and other modifiers. This is primarily the case for the sentence in Figure 4a. By contrast, *left-branching* sentences place verbs, objects, or modifiers before the main subject of a sentence [63]. Left-branching sentences are often cognitively more taxing as they involve more complex constructions that require a speaker to remember more information about the subject before the subject is explicitly mentioned. As a result, in English, the degree of left-branching within a particular parsing of a sentence can be used as a proxy for syntactic complexity.

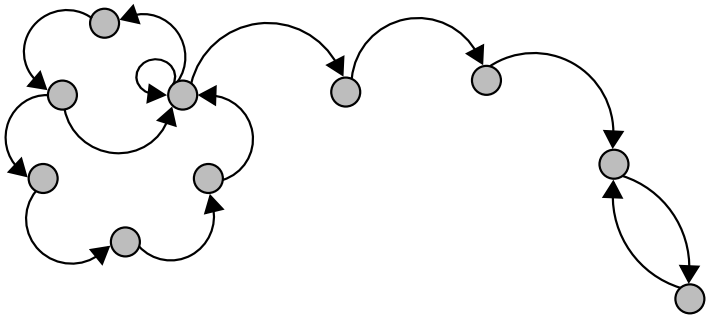
Once a parsing method has been implemented, various measures of lexical and syntactical complexity can be computed for each sentence or phrase. Yngve proposes one such method in [61]. Given the right-branching nature of simple English sentences, he proposes a measure of complexity based on the amount of left-branching in a given sentence. At each node in the parse tree, the rightmost branch is given a *score* of 0. Then, each branch to the left of it is given a score that is incremented by 1 when moving from right to left at a given node. The score for each token is the sum of scores up all branches to the root of the tree. An alternative scoring scheme for the same parse tree structure was proposed by Frazier [62]. He notes that *embedded clauses* within a sentence are an additional modifier that can increase the complexity of the syntactical construction of that sentence. Therefore, just as with left-branching language, the speaker or listener would need to retain more information in order to properly convey or interpret the full sentence, respectively. Frazier's scoring method emphasizes the use of embedded clauses when evaluating the syntactic complexity. The scores are assigned to each lexeme as in Yngve's scoring, but they are summed up

to the root of the tree or the lowest node that is not the leftmost child of its parent node. Examples of both Yngve and Frazier scoring can be seen in Figure 4a.

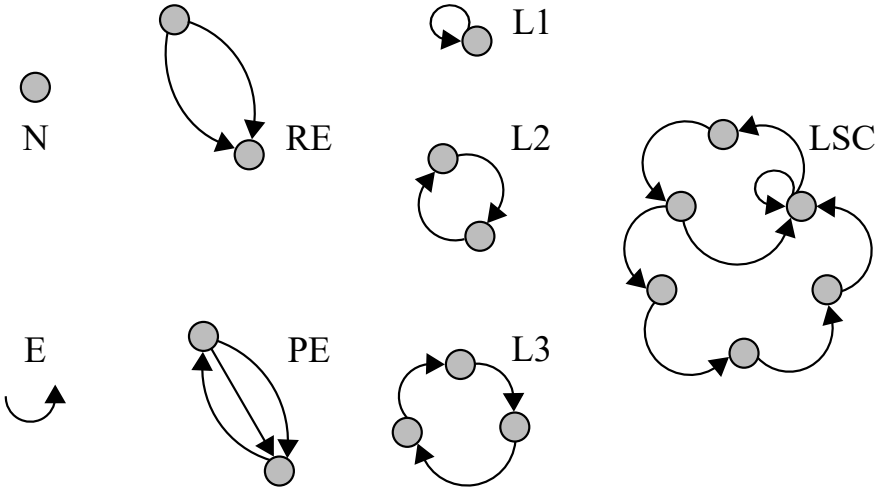
Another type of syntactical parsing of a sentence is known as *dependency parsing*, in which all nodes are treated as terminal nodes (no phrase categories such as *verb phrase* or *noun phrase*) [64]. A dependency-based parse tree aims to map the dependency of each word in a sentence or phrase to another within the same utterance. Methods proposed by Lin [65] and Gibson [66] provide some ways by which the lexical dependency distances can be determined. The general idea behind these methods is that longer lexical dependency distances within a sentence indicate a more complex linguistic structure, as the speaker and listener must remember more information about the dependencies between words in a sentence. An example of the same sentence is shown with a dependency-based parse tree in Figure 4b, also adapted from [49].

Mota *et al.* also propose a graph-theoretic approach for analyzing language structure as a marker of cognitive-linguistic ability with the construction of *speech graphs* [19, 67]. In this representation, the nodes are words that are connected to consecutive nodes in the sample text by edges representing lexical, grammatical, or semantic relationships between words in the text. As an examples, for a speech graph based on words in an utterance, the spoken language is first transcribed and tokenized into individual lexemes, with each unique word by a graph node. Directed edges then connect consecutive words[10]. The researchers in this work suggest that structural graph features, *i.e.* loop density, distance (number of nodes) between words of interest, *etc.*) serve as clinically relevant objective language measures that give insight into cognitive

[10]Speech graphs in some studies, *i.e.* [68], may use POS tags or other node structures



(a) Sample speech-graph representation of a spoken utterance. Each of the circular nodes represents a lexical unit (*e.g.* a single word) and the curved arrows represent edges which connect the relevant lexemes in the utterance. Attributes can be computed using the graph.



(b) Examples of *speech graph attributes* (SGAs). Examples include Nodes (N), Edges (E), Repeated Edges (RE) in same direction, Parallel Edges (PE), loops with 1, 2 and 3 nodes (L1, L2, L3), and the *largest strongly connected component* (LSC), *i.e.* the portion of the total graph that can be reached from all others when considering the directionality of edges.

Figure 5: (a) A sample speech-graph for a complete spoken utterance. (b) Example speech-graph attributes (SGAs). Both adapted from [67].

function. An example speech-graph representation structure of an arbitrary utterance is seen in Figure 5a. The computed *speech graph attributes* (SGAs) are the features which are extracted from these graphs, and some common ones visualized in Figure 5b.

The SGAs provide indirect measures of lexical diversity and syntactic complexity. For example, N is the number of unique words, E is the total number of words, and repeated edges represent repeated words or phrases in text.

2.2.3.2 Clinical Applications

The structural aspects of spoken language have been shown to have clinical relevance for understanding medical conditions that affect cognitive-linguistic ability. The previously mentioned work by Roark *et al.* also utilized several of the aforementioned methods to analyze the language of individuals with MCI and healthy control participants [49]. In addition to the lexeme-level features described in Section 2.2.2, they also considered Yngve [61] and Frazier [62] scoring measures from constituency-based parsing of the transcripts of participant responses[11]. Mean, maximum, and average Yngve and Frazier scores were computed for each participant's language samples. Roark *et al.* also used dependency parsing and computed lexical dependency distances, similar to the example in Figure 4b. Along with the lexical features and speech features, participants with MCI and healthy elderly control participants were classified successfully, as previously described in Section 2.2.2.

The speech-graph approach is used by Mota *et al.* to study the language of patients with schizophrenia and bipolar disorder (mania) [19, 67]. The researchers were able to identify structural features of the generated graphs (such as loop density, distance

[11] Using the Charniak parser [69]

between words of interest, *etc.*) that serve as objective language measures containing clinically relevant information (*e.g.* flight of thoughts, poverty of speech, *etc.*). Using these features, the researchers were able to visualize and quantify concepts such as the *logorrhea* (excessive wordiness and incoherence) associated with mania, evidenced by denser networks. Similarly, the *alogia* (poverty of speech) typical of schizophrenia was also visible in the generated speech-graph networks, as evidenced by a greater number of nodes per word and average total degree per node. Control participants, participants with schizophrenia, and participants with mania were classified with over 90% accuracy, significantly improving over traditional clinical measures, such as the Positive and Negative Syndrome Scale (PANSS) and Brief Psychiatric Rating Scale (BPRS) [19].

2.2.3.3 Advantages & Disadvantages

Consideration of sentence-level syntactical complexity offers several advantages that address some of the drawbacks of lexeme-level analysis. As the work discussed here reveals, sentence structure metrics via syntactic parsing or speech-graph analysis offer powerful information in distinguishing healthy and clinical participants with schizophrenia, bipolar disorder/mania, mild cognitive impairment, and potentially several other conditions. Since sentence construction further taxes the cognitive-linguistic system beyond word finding, methods that capture sentence complexity provide more insight into the neurological health of the individual producing these utterances. This provides a multi-dimensional representation of cognitive-linguistics and allows for better characterization of different clinical conditions, as Mota *et al.* did with patients with schizophrenia and those with bipolar disorder/mania [19].

However, while offering the ability to analyze more complex sentence structures, sentence-level syntactical analysis is also prone to increased complexity due to large range of implementation methodologies. For example, there are countless methods developed over the years for parsing language with different tools for measuring complexity relying on different algorithmic implementations of the language parsers, a widely studied topic in linguistic theory. A thorough empirical evaluation of the various parsing methods is required to better characterize the performance of these methods in the context of clinical applications.

2.2.4 Semantic Analysis

Cognitive function is also characterized by one's ability to convey organized and coherent thoughts through spoken or written language. Here, we will cover some of the fundamental methods in NLP for measuring semantic coherence that have been used in clinical applications.

2.2.4.1 Methods

Semantic similarity in natural language is typically measured computationally by *embedding* text into a high-dimensional vector space that represents its semantic content. Then, a notion of distance between vectors can be used to quantify semantic similarity or difference between the words or sentences represented by the vector embeddings.

Word embeddings are motivated by the *distributional hypothesis* in linguistics, a concept proposed by English linguist John R. Firth who famously stated "You

shall know a word by the company it keeps" [70], *i.e.* that the inherent meaning of words is derived from their contextual usage in natural language. One of the earliest developed word embedding methods is *latent semantic analysis* (LSA) [11], in which word embeddings are determined by co-occurrence. In LSA, each unit of text (such as a sentence, paragraph, document, *etc.*) within a corpus is modeled as a bag of words.

As per Firth's hypothesis, the principal assumption of LSA is that words which occur together within a group of words will be semantically similar. As seen in Figure 6, a matrix (A) is generated in which each row is a unique word in the text ($w_1, \ldots, w_n$) and each column represents a document or collection of text as described above ($d_1, \ldots d_d$). The matrix entry values simply consist of the count of co-occurrence statistics, that is the number of times each word appears in each document. Then a *singular value decomposition* (SVD) is performed on A, such that $A = U\Sigma V^{\mathrm{T}}$. Here, U and V are orthogonal matrices consisting of the left-singular and right-singular vectors (respectively) and Σ is a rectangular diagonal matrix of singular values. The diagonal elements of Σ can be thought to represent semantic categories, the matrix U represents a mapping from the words to the categories, and the matrix V represents a mapping of documents to the same categories. A subset of the r most significant singular values is typically chosen, as shown by the matrix $\hat{\Sigma}$ in Figure 6. This determines the dimension of the desired word embeddings (typically in the range of ~100-500). Similarly, the first r columns of U form the matrix $\hat{U}$ and the first r rows of V^{T} form the matrix $\hat{V}^{\mathrm{T}}$. The r-dimensional word embeddings for the n unique words in the corpus are given by the resulting rows of the product $\hat{U}\hat{\Sigma}$. Similarly, r-dimensional document embeddings can be generated by taking the d columns of the product $\hat{\Sigma}\hat{V}^{\mathrm{T}}$.

In recent years, several new word embedding methods based on neural networks have gained popularity, such as *word2vec* [14] or *GloVe* [15], which have shown

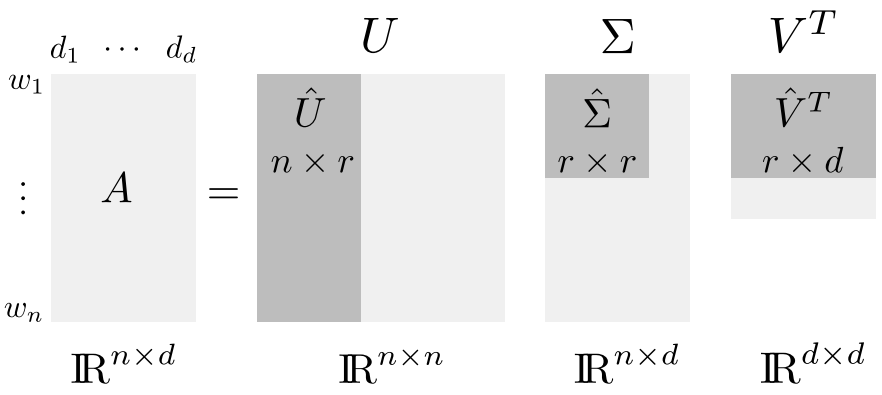


Figure 6: A visual representation of latent semantic analysis (LSA) by singular value decomposition (SVD).

improved performance over LSA for semantic modeling when sufficient training data is available [71]. As an example, we take a more detailed look at word2vec, proposed by Mikolov *et al.* , in which they present an efficient method for predicting word vectors based on very large corpora of text. They present two versions of the word2vec algorithm, a *continuous bag-of-words* (CBOW) model and *continuous skip-gram* model, seen in Figure 7. At the input in both implementations, every word in a corpus of text is uniquely *one-hot encoded*; *i.e.* in a corpus of V unique words, each word is uniquely encoded as a V-dimensional vector in which all elements are 0 except for a single 1. In both models, the inputs, $\mathbf{x} \in \mathbb{R}^V$, are multiplied by a weight matrix, $W \in \mathbb{R}^{V \times N}$ to obtain a hidden latent representation, $\mathbf{h} = W^{\mathrm{T}}\mathbf{x} \in \mathbb{R}^N$, with $N < V$ typically. The hidden representation is then multiplied by another weight matrix, $\tilde{W} \in \mathbb{R}^{N \times V}$ to obtain an output representation $\mathbf{u} = \tilde{W}^{\mathrm{T}}\mathbf{h} \in \mathbb{R}^V$. The softmax operation, given in Equation (2.5), is then performed on the elements u_j, $j = 1, \ldots, V$ of $\mathbf{u}$ to obtain an

output vector, $\mathbf{y}$, which approximates a one-hot encoded output prediction.

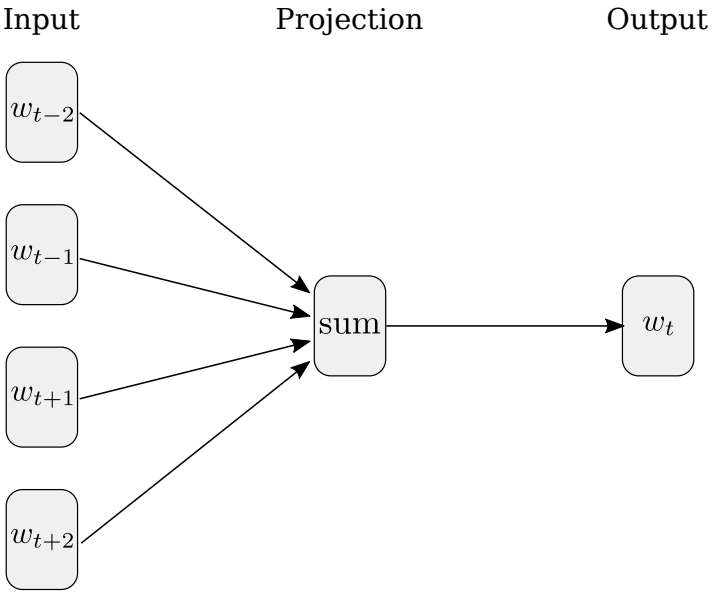
$$\mathbf{y} = \text{softmax}(u_j) = \frac{\exp u_j}{\sum_{i=1}^{V} \exp u_i}, \quad \mathbf{u} = [u_1, \ldots, u_V]^{\mathrm{T}} \tag{2.5}$$

In the CBOW implementation (Figure 7a), the inputs are the context words in the particular neighborhood of a target center word, w_t. In the skip-gram implementation (Figure 7b), the input is the center word and the objective is to predict the context words at the output. In both models, the latent hidden representation of dimension N gives an embedding for the word represented by the one-hot encoded input word. The training objective is to minimize the cross-entropy loss for the prediction outcomes.

There are several other methods for word embeddings, each relying on the distributional hypothesis and each with various advantages and disadvantages. For example, LSA, *word2vec* and *GloVe* are simple to train and effective, but a major disadvantage is that they do not handle out-of-vocabulary (OOV) words or consider words with multiple unrelated meanings. For example, the English word "bark" can refer to the bark of a dog or to the bark of a tree, but its vector representation would be an average representation, despite the drastically different usage in each context. Some methods based on deep neural networks (DNNs), such as recurrent neural network (RNN) / long-short term memory (LSTM) networks (*e.g.* ELMo [72]) or transformer architectures (*e.g.* BERT [73]) utilize contextual information to generate embeddings for OOV words.

In addition to individual words, embeddings can also be learned at the sentence level. The simplest forms of sentence embeddings involve unweighted averaging of LSA, *word2vec*, *GloVe*, or other embeddings. Weighted averages can also be computed, such as by using *term frequency-inverse document frequency* (tf-idf) generated weights or *Smooth Inverse Frequency* (SIF) [74]. Others have found success learning sentence

(a) Continuous bag-of-words (CBOW)



(b) Continuous skip-gram

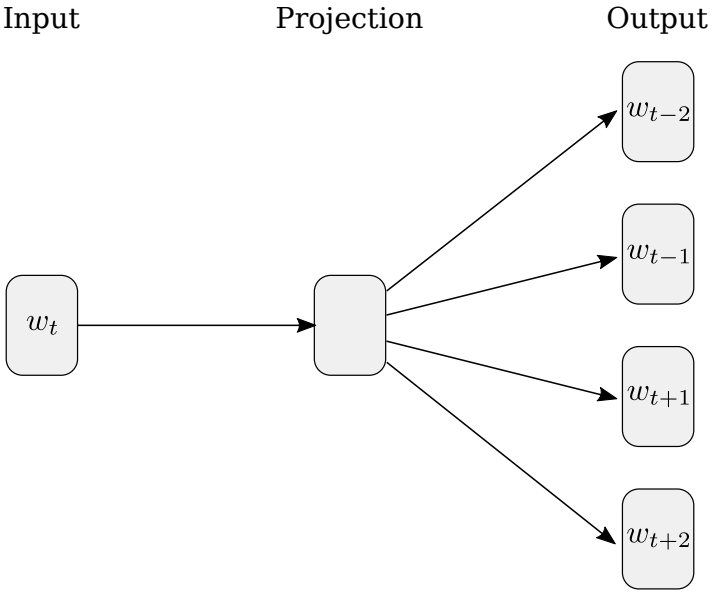


Figure 7: *word2vec* model architectures proposed in [14]. (a) In the CBOW model, the context words are inputs used to predict the center word. (b) In the skip-gram model, the center word is used to predict the context words.

representations directly, such as in *sent2vec* [75]. Whole sentence encoders, such as *InferSent* [76] and the *Universal Sentence Encoder* (USE) [77] offer the advantage of learning a full sentence encoding that considers word order within a sentence; *e.g.* the sentences "The man bites the dog" and the "The dog bites the man" will each have different encodings though they contain the same words.

Once an embedding has been defined, a notion of semantic similarity or difference must also be defined. Several notions of distance can be computed for vectors in high-dimensional space, such as Manhattan distance (ℓ_1 norm), Euclidean distance (ℓ_2 norm), or many others. Empirically, the *cosine similarity* (cosine of the angle, θ, between vectors) has been found to work well in defining semantic similarity between word and sentence vectors of many types. Cosine similarity can be computed using Equation (2.6) for vectors $\mathbf{w}_1$ and $\mathbf{w}_2$.

$$\text{CosSim}(\mathbf{w}_1, \mathbf{w}_2) = \cos\theta = \frac{\mathbf{w}_1^{\mathrm{T}}\mathbf{w}_2}{\|\mathbf{w}_1\|_2 \|\mathbf{w}_2\|_2} \tag{2.6}$$

In addition to word and sentence embedding semantic similarity measures, techniques such as *topic modeling* and *semantic role labeling* have also gained recently popularity in NLP and its applications to clinical language samples. *Latent dirichlet analysis* (LDA) is one such statistical topic modeling method which can be used to identify overarching themes in samples of text [78]. Other studies have utilized *semantic role labeling*, a probabilistic technique which automatically attempts to identify the semantic role a particularly entity plays in a sentence [79].

2.2.4.2 Clinical Application

Many forms of mental illness can result in a condition known as *formal thought disorder* (FTD), which impairs an individual's ability to produce semantically coherent

language. FTD is most commonly associated with schizophrenia but is often present in other forms of mental illness such as mania, depression, and several others [80, 81]. Some common symptoms include poverty of speech (*alogia*), derailment of speech, and semantically incoherent speech (*word salad*) [81, 82]. Language metrics that track semantic coherence are potentially useful in clinical applications, such as measuring the coherence of language as it relates to FTD in schizophrenia. One of the first studies to demonstrate this was conducted by Elvevåg *et al.* [10]. The language of patients with varying degrees of FTD (rated by standard clinical scales) was compared with a group of healthy control participants. The experimental tasks consisted of single word associations, verbal fluency (naming as many words as possible within a specific category), long interview responses (~1-2 minutes per response), and storytelling. LSA was utilized to embed the word tokens in the transcripts. The semantic coherence in each task was computed as follows:

- *Word Associations*: Cosine similarity between cue word and response word, with an average coherence score for each participant
- *Verbal Fluency*: Cosine similarity between first and second word, second and third word, etc. were computed, with an average coherence score computed per participant
- *Interviews*: Cosine similarity was computed between the question and participant responses. An average word vector was computed for the prompt question from the interviewer. Then a *moving window* (of size 2-6 words) for the participant response was used to average all the word vectors within the window and compute a cosine similarity between the question and response. The window was moved over the entire participant response and a new cosine similarity was computed between the question and response window until reaching the end

of the response. This method tracks how the cosine similarity behaves as the participant response goes farther from the question, with the expectation that the response would be more tangential over time with decreased coherence as the participant moves farther from the question. A regression line was fit for each participant to measure the change in cosine similarity coherence over time, and the slope of the line was computed to measure the tangentiality of the response per participant.

- *Storytelling*: Cosine similarity of the participant's response was compared to the centroid participant response for all narrative utterances of the same story. This was used to predict the clinical rating for thought disordered language samples when asked to tell the same story.

They demonstrated that the control participants had higher coherence scores compared to the FTD groups across all tasks.

In a more recent study, predictive features of language for the onset of psychosis were studied by Bedi *et al.* [12]. Open-ended narrative-like interview transcripts of young individuals who were determined to be at *clinical high-risk* (CHR) for psychosis were collected and analyzed to predict which individuals would eventually develop psychosis. Participants were tracked and interviewed over a period of two and a half years. In this study, LSA was again used to generate word embeddings. An average vector for each phrase was computed, and a cosine-similarity measure was computed to measure the semantic coherence between consecutive phrases (*first-order coherence*) and every other phrase (*second-order coherence*).

A distribution of the first and second-order coherence scores (cosine similarities) was compiled for each participant, and several statistics were computed based on the distribution of coherence scores, *e.g.* maximum, minimum, standard deviation, mean,

median, 10^{th} percentile, and 90^{th} percentile. Each of these statistics was considered as a separating feature between the clinical and control samples. In addition to the semantic analysis, POS-tagging was performed to compute the frequency of use of each part-of-speech to obtain information about the structure of each participant's naturally-produced language. The language features with the best predictive power in the classifier were the minimum coherence between consecutive phrases for each participant (maximum discontinuity) and the frequency of use of determiners (normalized by sentence length). This initial study only had 34 participants total (only 5 CHR+ participants) and was intended as a proof-of-principle exploration. In an expansion of this work, Corcoran *et al.* trained their classifier using two larger datasets, in which one group of participants was questioned with a prompt-based protocol and another group of participants was given a narrative protocol in which they were required to provide longer answers (similar to the previous work) [13]. They note that the first and second-order coherence metrics collected in the previous study were useful for determining semantic coherence with the narrative-style interview transcripts with longer responses. However, for the shorter prompt-based responses (often under 20 words), it is often difficult to obtain these metrics. Therefore, coherence was-computed on the *word-level* rather than phrase-level by computing the cosine similarity between word embeddings within a response with an inter-word distance of k, with k ranging from 5 to 8. As before, typical statistics were computed on the coherence values obtained for each participant response (maximum, minimum, mean, median, 90^{th} percentile, 10^{th} percentile, *etc.*). They were able to successfully predict the onset of schizophrenia by discriminating the speech of healthy controls and those with early onset schizophrenia with $\sim 80\%$ accuracy.

Other studies make use of a variety of linguistic features to predict the presence of

clinical conditions. For example, Kayi *et al.* identified predictive linguistic features of schizophrenia by analyzing laboratory writing samples of patients and controls for their semantic, syntactic, and pragmatic (sentimental) content [20]. A second dataset of social media messages from self-reporting individuals with schizophrenia over the Twitter API was also evaluated for the same types of content. The semantic content of the language was quantified by three methods: First, semantic role labeling was performed using the Semafor tool [79] to identify the role of individual words within a sentence or phrase Then, LDA was used to identify overarching themes that separated the clinical and control writing samples [78]. LDA identifies topics in the text and also identifies the top vocabulary used in each topic. Finally, clusters of word embeddings within the writing were generated using the k-means algorithm and *GloVe* word vector embeddings [15]. The frequency of each cluster was computed per document by checking the use of each word of the document in each cluster. The syntactic features used in this study again were obtained by computing the frequency of use of parts of speech (found by POS tagging) and by generating parse trees, using tools optimized for the corpus. Lastly, pragmatic features were found by performing *sentiment analysis* to classify the sentiment of the writing samples into distinct groups (*very negative*, *negative*, *neutral*, *positive*, *very positive*). They successfully showed a distinct set of predictive features that could accurately separate participants with schizophrenia from healthy controls in all of the language analysis categories. However, when using a combination of features and various machine learning classifiers (random forest and support vector machine), they found that utilizing a combination of the semantic and pragmatic features led to the most promising accuracy (81.7%) in classification of control participants and those with schizophrenia. The limited availability of language data in schizophrenia is always a difficult challenge, so another

study by Mitchell *et al.* analyzed publicly available social media (Twitter) posts by self-identifying individuals with schizophrenia using LDA, LIWC generated-vectors, and various clustering techniques to show statistically significant differences in their language patterns when compared to general users [83].

Another vector-space topic modeling approach was developed by Yancheva and Rudzicz for analyzing transcripts of picture description tasks for participants with AD and healthy controls [84]. They propose a general method for generating *information content units* (ICUs), or topics of interest, from common images used in clinical description task evaluations, *i.e.* the famous *Cookie Theft* picture with reference speech samples [85]. The generated ICUs were compared with human-supplied ICUs from common usage in clinical practice, and most of the categories exhibited a close match. The study found that participants with AD and healthy controls were likely to discuss the same topics, but those with AD had wider topic clusters with more irrelevant directions. Additionally, they were able to find a small set of generated ICUs that had slightly better classification performance than a much larger set of human selected ICUs for the same task, with $\sim 80\%$ accuracy. Related work by Hernández-Domínguez *et al.* took a similar approach to generate a population-specific set of categories for participants with AD ($n = 257$), MCI ($n = 43$), and healthy controls ($n = 217$) [86]. The resulting features were significantly correlated with severity as assessed by the MMSE, and classification performance was characterized by the receiver operating characteristic (ROC) area under curve (AUC) performance of AUC ≈ 0.76 for all three groups.

2.2.4.3 Advantages and Disadvantages

While these studies have been successful in measuring the semantic coherence of language as it relates to thought disorders, there are several limitations. Recent work by Iter *et al.* identifies and attempts to address some of these shortcomings when measuring semantic coherence for FTD in schizophrenia [16]. Interviews with a small sample of patients were collected and just the participant responses (of ~300 words each) were analyzed for their semantic content. They noted that when using the *tangentiality model* of semantic coherence (i.e. regression of the coherence over time with the sliding window) of Elvevåg *et al.* [10] and the *incoherence model* of semantic coherence of Bedi *et al.* [12], they were unable to convincingly separate their clinical and control participants based on language analysis. One reason for this was due to the presence of verbal fillers, such as "um" or "uh" and many *stop words* without meaningful semantic content. Another reason is that longer sentences (or long moving windows) tend to be scored as more coherent due to a larger overlap of words. The third reason they identified (but did not address) is that repetitive sentences and phrases would be scored as highly coherent, even though repetition of ideas is common in FTD and should be scored negatively. The authors proposed a series of improvements to address some of these limitations. However, the sample sizes in this study were small (9 clinical participants and 5 control participants).

2.3 Measuring Cognitive and Thought Disorders with Speech Signal Processing

While cognitive-linguistic health is more directly observed through analysis of complex language production, additional information can be derived by speech signal

analysis of individuals with cognitive impairments or thought disorders. This is because the acoustic stream is the physical manifestation of the cognitive-linguistic processing that has gone into creating the message being conveyed, in near real-time. In this way, pauses during speech can be associated with difficulty in lexical retrieval (word-finding difficulties) or with extra processing time needed for message formulation. Pressed speech, that which is rapidly produced without insertions of natural pauses, can be associated with mania and "flight of thoughts". Conversely, reductions in the rhythmic and melodic variations in speech may be indicative of changes in mood.

The information derived from the speech signal is used alone or in conjunction with many of the previously described methods to assess cognitive-linguistic health. This is either done directly by measuring different aspects of speech production including prosody, articulation, or vocal quality; or is done as a pre-processing step by using automatic speech recognition (ASR) for transcription of speech samples for follow-on linguistic analysis.

In this section, we will review how various signal processing methods are used to extract clinically-relevant insight from an individual's speech samples for additional insight into detection of disorders that affect cognition and thought. Referring back to Fig. 2, these include features extracted from vocal fold vibration (source), movement of the articulators (filter), and the overall rhythm of the speech signal (prosody).

2.3.1 Methods

2.3.1.1 Prosodic Features

Prosody refers to the rhythm and melody of speech. Examples of computable temporal prosodic features from recorded speech signals include the duration of voiced segments, duration of silent segments, loudness, measures of periodicity, fundamental frequency (F_0), and many other similar features [49, 87]. These measures can indicate irregularities in the rhythm and timing of speech. Additionally, nonverbal speech cues, *e.g.* counting the number of interruptions, interjections, natural turns, and response times can also indicate identifying features of irregular speech patterns [88].

2.3.1.2 Articulation Features

Several spectral features that capture movement of the articulators have been used in the clinical speech literature to measure the acoustic manifestation of the cognitive-linguistic deficits discussed in Section 2. These include computing statistics related to the presence of additional formant harmonic frequencies, *i.e.* F_1, F_2, and F_3, computing formant trajectories over time [89], or computing the vowel space area [90]. The *spectral centroid* can also be computed for each frame of speech signal that is analyzed [91]. The spectral centroid is essentially the center of mass for the frequency spectrum of a signal, and relates to the "brightness" or timbre of the perceived sound for audio.

Time-frequency signal processing techniques are also commonly used since acoustic speech signals are highly non-stationary. For example, computation of the mel-

frequency cepstral coefficients (MFCC) with the mel scale filterbank provides a compressed and whitened spectral representation of the speech [92]. These features are often used as inputs into an automatic speech recognition (ASR) system, but can also be monitored over time to identify irregularities in speech due to cognitive or thought disorders. As an example, common statistical features such as the mean, variance, skewness, and kurtosis of the MFCCs over time can be tracked for identification of irregularities between healthy individuals and those with some cognitive or thought disorders [50].

2.3.1.3 Vocal Quality Features

There is evidence that there are vocal quality changes associated with cognitive disorders [93]. These can be measured from the speech signal by isolating the *source* of speech production, involving the flow of air through the lungs and glottis and affecting perceptible voice quality. Voice quality measures that have previously been used in the context of cognitive and thought disorders include:

- *jitter*: small variations in glottal pulse *timing* during voiced speech
- *shimmer*: small variations in glottal pulse *amplitude* during voiced speech
- *harmonic-to-noise ratio* (HNR): the ratio of formant harmonics to inharmonic spectral peaks, *i.e.* those that are not whole number multiples of F_0

These features alone are often difficult to consistently compute and interpret, but can provide insight for the diagnosis and characterization of certain clinical conditions.

2.3.1.4 Automatic Speech Recognition

Recent improvements in ASR and in tools for easily implementing ASR systems have made possible the use of these systems in clinical speech analysis. This is most commonly done by using ASR in place of manual transcription for the extraction of linguistic features (*i.e.* features covered in Section 2.2); however, this is often more error prone with regard to incorrect word substitutions, unintended insertions, or unintended deletions in the automatically generated transcript. The *word error rate* (WER) for an utterance of N words is given in Equation (2.7),

$$\text{WER} = \frac{\#\text{ of insertions} + \text{deletions} + \text{substitutions}}{N}, \tag{2.7}$$

and is a typical statistic used to evaluate the performance of an ASR system. It is often more difficult to maintain high accuracy (low WER) for ASR with pathological speech samples, as the relative dearth of this data makes it difficult to train reliable ASR models optimized for this task. Other studies have also made use of ASR for paralinguistic feature extraction, such as the automated detection of filled pauses, natural turns, interjections, *etc.* Understanding the effects of ASR errors on downstream NLP tasks is an important area to address in which the current work is limited. Some recent attempts have been made to simulate ASR errors on text datasets and evaluate their effects on downstream tasks [94, 95, 24]. These potentially have future applications in language models that can analyze noisy datasets with ASR errors in clinical practice.

2.3.2 Clinical Applications

2.3.2.1 Acoustic Analysis

Disorders such as PPA, MCI, AD, and other forms of dementia are associated with a general slowing of thoughts in affected individuals. This has been shown to have detectable effects on speech production through acoustic analysis. In a study by König *et al.* , healthy controls and participants with MCI and AD were recorded as they were asked to perform various tasks, such as counting backwards, image description, sentence repeating, and verbal fluency testing [87]. Temporal prosodic features such as the duration of voiced segments, silent segments, periodic segments, and aperiodic segments were all computed. Then, the ratio of the mean durations of voiced segments to silent segments were also computed as features to express the continuity of speech in the study's participants. As expected, it was shown that healthy control participants showed greater continuity in these metrics when compared to those with MCI or AD. These quantifiable alterations of speech in individuals with MCI and AD allowed the researchers to successfully separate patients with AD from healthy controls (approx. 87% accuracy), patients with MCI from controls (approx. 80% accuracy), and patients with MCI from patients with AD (approx. 80% accuracy). López-de-Ipiña *et al.* conducted another study in which acoustic features (related to prosody, spectral analysis, and features with emotional content) were extracted from spontaneous speech samples to classify participants with AD at different stages (early, intermediate, and advanced) [96]. Among the computed prosodic features were the mean, median, and variance for durations of voiced and voiceless segments. Short-time energy computations were also computed for the collected samples in the time-domain. In the frequency-domain, the spectral centroid was determined for each speech sample. The authors also claim that features such as the contour

of F_0 and source features like shimmer, jitter, and noise-harmonics ratio contain emotional content that can be useful in the automatic AD diagnosis. Lastly, they propose a new feature, which they term *emotional temperature* (ET), which is a normalized (independent of participant) measure ranging from 0-100 based on several of prosodic and paralanguistic features that were previously mentioned[12]. The study revealed several interesting findings. First, the spontaneous speech analysis indicated that participants with AD exhibited higher proportions of voiceless speech and lower proportions of voiced speech, indicating a loss of fluency and shortening of fluent speech segments for those with AD. While classification accuracy was good when using a set of prosodic speech features, they noted that accuracy improved when the emotional features (*i.e.* the proposed ET metric) were used[13].

Acoustic analysis of speech can make use of ASR to count dysfluencies in spoken language that are often associated with neurodegenerative decline. Pakhomov *et al.* made an early attempt to use ASR to extract many such prosodic features (pause-to-word ratio, normalized pause length, *etc.*) on picture-description task transcripts for participants with three variants of Frontotemporal Lobar Degeneration (FTLD) [97]. A more recent pair of studies by Tóth *et al.* explored using ASR for detection of MCI [98, 99]. However, in their work, only acoustic features were considered, and precise word transcripts were not required, mitigating the effect of the typically high WER for clinical speech samples. Instead, the authors trained a new ASR model with a focus of detecting individual phonemes. The features considered in this study were mostly prosodic (articulation rate, speech tempo, length of utterance, duration

[12]The example in [96] shows that a typical ET value is approx. 95 for healthy control participants and approx. 50 for those with AD

[13]see Figure 9 in [96]

of silent and filled pauses, the number of silent and filled pauses). The focus of the study was to compare the effects of manually annotating transcripts with the faster ASR method. Since most ASR models cannot differentiate between filled pauses and meaningful voiced speech, their detection was a major focus of this work for automated MCI detection. The ASR model was trained with annotated filled pause samples to learn to detect them in spontaneous speech. The authors were able to show comparable results between the ASR and manual methods for MCI detection with the same feature set (82% accuracy for manual vs. 78% for ASR) [98].

While acoustic speech processing on its own has been less explored in detecting thought-disorder related mental illness, some researchers have found ways in which useful information can be derived solely from speech signals for this purpose. One example is seen in work by Tahir *et al.* [88]. In this study, patients with severe schizophrenia, receiving Cognitive Remediation Therapy (CRT), were differentiated from control participants with less severe schizophrenia (no CRT recommended) by *non-verbal* speech analysis. They note that nonverbal and conversational cues in speech often play a crucial role in communication, and that it is expected that individuals with schizophrenia would have a muted display of these features of speech. Cues used as inputs to a classifier included interruptions, interjections, natural turns, response time, speaking rate, among others. Preliminary results from this study with participants with severe schizophrenia ($n = 8$) and less-severe forms of the disease ($n = 7$) indicate that these nonverbal cues show approximately 90% accuracy in classifying control participants from those with more severe forms of schizophrenia. They also attempted to validate the computed features by examining their correlation with traditional subjective clinical assessments. Some of the computed objective nonverbal speech cue features had high correlation with subjective assessments; *e.g.* "poor rapport

with interviewer" has a strong correlation with longer participant response times. The acoustics of bipolar disorder have also been studied, for example by Guidi *et al.* [100]. In this study, the authors propose an automated method for estimating the contour of F_0 over time with a moving window approach as a proxy for mood changes. In particular, they study local rising and falling events of the F_0 contour, including positive and negative slopes, amplitude, duration, and tilt to indicate different emotional states. The features were first validated on a standard emotional speech database and then used to classify bipolar patients ($n = 11$) and healthy control subjects ($n = 18$). They noted that intra-subject analysis showed good specificity in classifying bipolar subjects and healthy controls across all contour features, but that directions of most were not consistent across different subjects. Due to limited data, they propose a study with a larger number of subjects including glottal, spectral, and energy features.

2.3.2.2 Combination of Acoustic and Textual Features

Many dementia studies also use both acoustic and textual data with promising results. As an example, the previously mentioned work by Roark *et al.* (in Section 2.2) also made use of acoustic speech samples to aid in the detection of MCI from naturally-produced spoken language. The researchers used manual and automated methods to estimate features related to the duration of speech during each utterance, including the quantity and duration of pause segments. Some of the features that were computed include fundamental frequency, total phonation time, total pause time, pauses per sample, total locution time (both phonation and pauses), verbal rate, and several others [49]. They conclude that automated speech analysis produces very similar

results to manually computing these metrics from the speech samples, demonstrating the potential of automated speech signal processing for detecting MCI. Additionally, they found that a combination of linguistic complexity metrics and speech duration metrics lead to improved classification results. The previously described work on PPA subtypes in [55] was expanded by Fraser *et al.* in [101]. Acoustic features were also extracted and added to the previous set of linguistic features to improve the classification results of PPA subtypes (PNFA and semantic dementia) and healthy control participants. The added acoustic features included temporal prosodic features (*i.e.* speech duration, pause duration,pause to word ratio, *etc.*), mean and variance of F_0 and first three formants (F_1, F_2, F_3), mean instantaneous power, mean and maximum first autocorrelation function, instantaneous power, and vocal quality features, *i.e.* jitter and shimmer. The authors tested the relative significance of all features using different feature reduction techniques and noted that more textual features were usually selected in each case. However, the addition of acoustic features had the greatest positive impact when attempting to differentiate between healthy control participants and those with one of the PPA subtypes, but proved less useful in distinguishing the subtypes. Their later study on AD [50] also used a similar hybrid approach with speech and language metrics to show good classification separating AD participants from healthy controls [50]. The DementiaBank[14] corpus was used to collect the data for this analysis. The study considered 370 distinctive features; linguistic features included grammatical features (from part-of-speech tagging), syntactic complexity (*e.g.* mean length of sentences, T-units, clauses, and maximum Yngve depth scoring for the parse tree, as described above), information content (specific and nonspecific word use), repetitiveness of meaningful words, and many more. Acoustic features associated with

[14]https://dementia.talkbank.org/access, accessed August 20, 2019

pathological speech were also identified by computation of MFCCs, their derivatives, and their second derivatives. To differentiate the clinical and control group, they considered mean, variance, skewness, and kurtosis of the MFCCs over time. After performing factor analysis on these features, they showed that most of the variance between controls and those with AD could be explained by semantic impairment, acoustic abnormalities, syntactic impairment, and information impairment.

2.3.2.3 Impact of ASR on Textual Features

Several studies have also used ASR to generate transcripts of spoken language tasks for textual feature extraction for dementia detection. However, unlike the phone-level ASR model built in [98] and [99], this use case does require accurate word-level transcripts (*i.e.* a low WER). Previous work has shown that ASR accuracy is reduced for both elderly patients and those with dementia [102, 103, 104]. To address this, Zhou *et al.* performed a study in which the DementiaBank[14] corpus was used to train an ASR model on domain data with elderly patients, both with and without AD [105]. They were able to show that an ASR model trained with a smaller in-domain dataset could improve WER-based accuracy than one trained with a larger out-of-domain dataset. Additionally, they were able to confirm that even with their model, diagnostic accuracy decreases with increasing WER, as expected, but the correlation between the two is relatively weak when selecting certain features that are more robust to ASR errors (such as word frequency and word length related features)[15].

Mirheidari *et al.* also used ASR with a combination of acoustic (temporal prosodic)

[15]The authors identify features that provide best diagnostic ability for gold-standard manual transcripts and transcripts with varying WER and ASR to identify these robust features, but they do not claim to understand why certain features seem more robust than others

and textual features (syntactical and semantic features) to diagnose and detect participants with neurodegenerative dementia (ND) and differentiate them from those with non-dementia related Functional Memory Disorder (FMD) with a conversational analysis dataset [106, 107]. With manual transcriptions, the classification accuracy was over 90% in classifying the two groups, but it dropped to 79% when ASR was used. As expected, they found that the significance of the syntactic and semantic textual features is diminished when transcriptions contain ASR errors. Sadeghian *et al.* attempted to improve the issue of transcription errors by training a custom ASR model using collected speech samples from participants with AD ($n = 26$) and healthy controls ($n = 46$) [108]. This was done by limiting the potential lexicon to the collected speech in their dataset as well as cleaning the audio files to reduce the WER. Their study used a combination of acoustic features (temporal prosodic features and F_0 statistics) and textual features computed from both manual and ASR-generated transcripts (POS tags, syntactic complexity measures from [49], idea density, and LIWC features [42]). In their work, the best classification results (over 90%) were seen when feature selection was performed using both the MMSE scores and computed acoustic and textual features, but using the computed features alone was nearly comparable and outperformed the MMSE scores on their own. Weiner *et al.* [109] instead compared the difference in the analysis of manual and ASR-derived transcripts for a large range of acoustic (prosody and timing related) and textual features (lexical diversity with Brunet's index and Honoré's statistic) for comparing participants with dementia and healthy controls. The off-the-shelf ASR model used in this work had a relatively high WER, but they were interestingly able to show that the WER itself was a reliable feature for classifying the different types of subjects. Additionally, many of the features they selected showed robustness to transcription

quality, possibly even taking advantage of the poor ASR performance to identify participants with dementia.

2.3.3 Advantages & Disadvantages

It is intuitive that fine-grained and discrete measures of "what is said" (language, in terms of lexical diversity, lexical density, semantic coherence, language complexity, *etc.*) may more directly capture early cognitive-linguistic changes in illness and disease than measures of "how it is said" (analysis of speech acoustics). However, emerging data shows that acoustic analysis offers converging and complementary information to several of the textual features discussed in Section 2.2. Most interestingly, changes in the outward flow of speech may precede measurable language-based changes [88, 101].

A particular advantage of evaluating speech acoustics is that ASR or transcription is not necessarily a required step. Automated acoustic metrics can be extracted from non-labeled speech samples [87, 88, 96, 110]. Further, some of the metrics provide complementary and interpretable value that cannot be gleaned from transcripts (rate, pause metrics, speech prosody). These directly correspond with subjectively described clinical characteristics (*e.g.* pressed speech, halting speech, flat affect etc). A disadvantage is that not all acoustic metrics offer that level of transparency. This is a running theme in clinical speech analysis. Many of these features are not currently used in clinical diagnosis despite their powerful predictive power because they are difficult to directly interpret (*e.g.* MFCCs); this means that clinicians can see the output of a complicated model but not understand why the model came to that decision or if it is considering clinically-relevant dimensions. For this reason, some effort has been undertaken to map the information contained in high-dimensional data

to be easily visualized and interpreted by clinicians, but this remains a significant challenge [111, 112].

2.4 Preliminary Results (Interspeech 2019 Paper)

Another issue with semantic coherence computation in clinical practice is difficulty with interpretability of computed metrics; for example, the cosine similarity between high dimensional word vectors is a somewhat abstract concept which is difficult for most to visualize. Our recent work [17] was a preliminary study that attempted to address this issue by computing semantic coherence measures (using *word2vec*, *InferSent*, and SIF embeddings), lexical density and diversity measures, and syntactic complexity measures as they relate to the language of patients with schizophrenia, patients with bipolar disorder, and healthy controls undergoing a validated clinical social skills assessment [21]. Linear regression was used to determine a subset of language features across all categories that could effectively model the scores assigned by clinicians during the social skills performance assessment, in which participants were required to act out various role-playing conversational scenes with clinical assessors scored for cognitive performance. Then, these features were used to train simple binary classifiers (both naïve Bayes and logistic regression), for which leave-one-out cross-validation was used to determine their effectiveness at classifying groups of interest. For classifying clinical (patients with schizophrenia and bipolar I disorder) participants and healthy control participants, the selected feature subset achieved ROC curve AUC performance of AUC ≈ 0.90; for classifying within the clinical group (to separate participants with schizophrenia and bipolar disorder), the classifier performance achieved AUC ≈ 0.80.

2.4.1 SSPA Data Collection

Our study involves the analysis of interview transcripts collected from a total of 87 clinical subjects and 22 healthy controls that participated in the SSPA task described by Patterson *et al.* [21]. Of the clinical population, 44 had been diagnosed with bipolar I disorder and 43 had been diagnosed with schizophrenia or schizoaffective disorder (considered together in this analysis). The SSPA interviews are described by Bowie *et al.* in [23]. The transcriptions used in our analysis were completed at Queen's University in Kingston, ON, Canada.

The task consists of three role-playing scenes: (1) 1-minute practice scene of making plans with a friend (not scored), (2) 3 minutes of greeting a new neighbor, and (3) 3 minutes of negotiation with a recalcitrant landlord over fixing an unrepaired leak. Each session was recorded and scored by trained research assistants upon reviewing the recording. Scene 2 (new neighbor) and Scene 3 (negotiation with landlord) were scored on a scale of 1 (low) to 5 (high) on several categories, *i.e.* interest/disinterest, fluency, clarity, social appropriateness, negotiation ability, *etc.* A composite score for each scene and an overall score is computed by averaging Scene 2 and Scene 3 scores.

Bowie *et al.* identified group differences between the scores of both clinical populations and healthy control subjects in [23] by evaluation on the SSPA task and several other clinical measures. In this work, we aim to automate this task with a subset of language metrics from the SSPA transcripts. Our first goal is to identify semantic and lexical features from which we can reliably predict SSPA performance. Then, we test the ability of these features to differentiate between healthy control and clinical populations, and we also test their ability to differentiate within the distinct groups in the clinical population.

2.4.2 Computed Language Features

In our work, we attempt to identify a comprehensive set of objective language measures from which we can model and predict SSPA performance and classify individuals using these features. Inspired by much of the previous work described in Section 2.2, we theorized that it is critical to consider language features that model semantic coherence through the use of word and sentence embeddings. We focused on a few pre-trained neural embedding models that are publicly available and known to model semantic similarity accurately. Additionally, we consider a set of lexical complexity features that are measures of lexical and syntactic complexity, described below.

2.4.2.1 Semantic Coherence

Many of the previously described studies in this area involve computing a notion of semantic coherence in language with the use of word embeddings in high-dimensional vector space, either with LSA or neural word embedding techniques [10, 12, 113, 16]. In nearly all cases, word and sentence/phrase embedding pairs, denoted by vectors $\mathbf{a}$ and $\mathbf{b}$, are evaluated with the notion of *cosine similarity*, a measure of the cosine of the angle $\theta_{\mathbf{a},\mathbf{b}}$ between the two vectors. We also use cosine similarity as a measure of pairwise sentence similarity, but with some modifications in implementation due the difference in the nature of the SSPA task and data collection.

Our work differs from several of the previously discussed studies in that we are interested in conversational semantic similarity between the subject and clinical assessor in each of the three scenes of the SSPA task. Therefore, we sought to utilize some of the

latest sentence/phrase embedding methods to compute a vector representation for each assessor and subject speaking turn. Then, we used the cosine similarity to compute the similarity score between each consecutive assessor + subject speaking turn, generating a distribution of similarity scores for each embedding method for each subject in each transcribed scene. The following sentence embedding representations are used in our analysis: (1) an unweighted bag-of-words (BoW) average for all word vectors based on the pre-trained *skip-gram* implementation of *word2vec* trained on the Google News corpus [14], (2) *Smooth Inverse Frequency* (SIF) with pre-trained skip-gram *word2vec* vectors [74], and (3) *InferSent* (INF) sentence encodings based on pre-trained *FastText* vectors [76]. The BoW average of vectors and SIF embeddings showed good baseline performance in [16], and we additionally included *InferSent*, a deep neural network sentence encoder, due to its strong performance on semantic similarity tasks. Then, basic statistics for the similarity score distribution were computed for each subject and transcribed scene. These included minimum, maximum, mean, median, 90^{th} percentile, and 10^{th} percentile coherence.

2.4.2.2 Linguistic Complexity

While semantic coherence measures are often the most effective at classifying patients with schizophrenia and bipolar disorder, several other linguistic complexity measures are used for a more holistic analysis. We consider a subset of these features, computed for the entire set of subject responses across all three scene transcripts.

Lexical diversity refers to unique vocabulary usage for a particular subject and for which several measurement techniques exist. The *type-to-token ratio* (TTR) is a well-known measure of lexical diversity, in which the number of unique words

(word *types*, V) are compared against the total number of words (word *tokens*, N): $\text{TTR} = {}^{V}/_{N}$. However, TTR is known to be negatively impacted for longer utterances, as the diversity of unique words plateaus as the number of total words increase. Hence, we consider a small selection of modified measures for lexical diversity in our work. The moving average type-to-token ratio (MATTR) [43] is one such method which aims to reduce the dependence on text length by considering TTR over a sliding window of the text. *Brunét's Index* (BI) [44], defined in Equation (2.2), is another measure of lexical diversity that has a weaker dependence on text length. A smaller value indicates a greater degree of lexical diversity. An alternative is also provided by *Honoré's Statistic* (HS) [45], defined in Equation 2.3, which emphasizes the use of words that are spoken only once (denoted by V_1). MATTR, BI, and HS have been used successfully in computational linguistics studies for patients with Alzheimer's disease [50, 54] and may prove to be similarly useful in our task.

Because we expect schizophrenia and bipolar patients to sometimes exhibit poverty of speech, we considered a few measures of lexical and syntactic complexity in our work.

Lexical density, which quantifies the degree of information packaging in a given text, is defined as the proportion of *content words* (*i.e.* nouns, verbs, adjectives, adverbs) [114]. Typically, these words convey more information than *function words*, *e.g.* prepositions, conjunctions, interjections, *etc.* We make use of the Stanford tagger [48] to compute POS tags to determine the number of function words (FUNC) and total words (W) and measure ${}^{\text{FUNC}}/_{\text{W}}$, which represents an inverse of the lexical density. A related, more granular measure is the proportion of interjections (UH) to the total words, which is given by ${}^{\text{UH}}/_{\text{W}}$. The mean length of sentence (MLS) is another easily computed measure which we expect to be lower for clinical subjects when compared

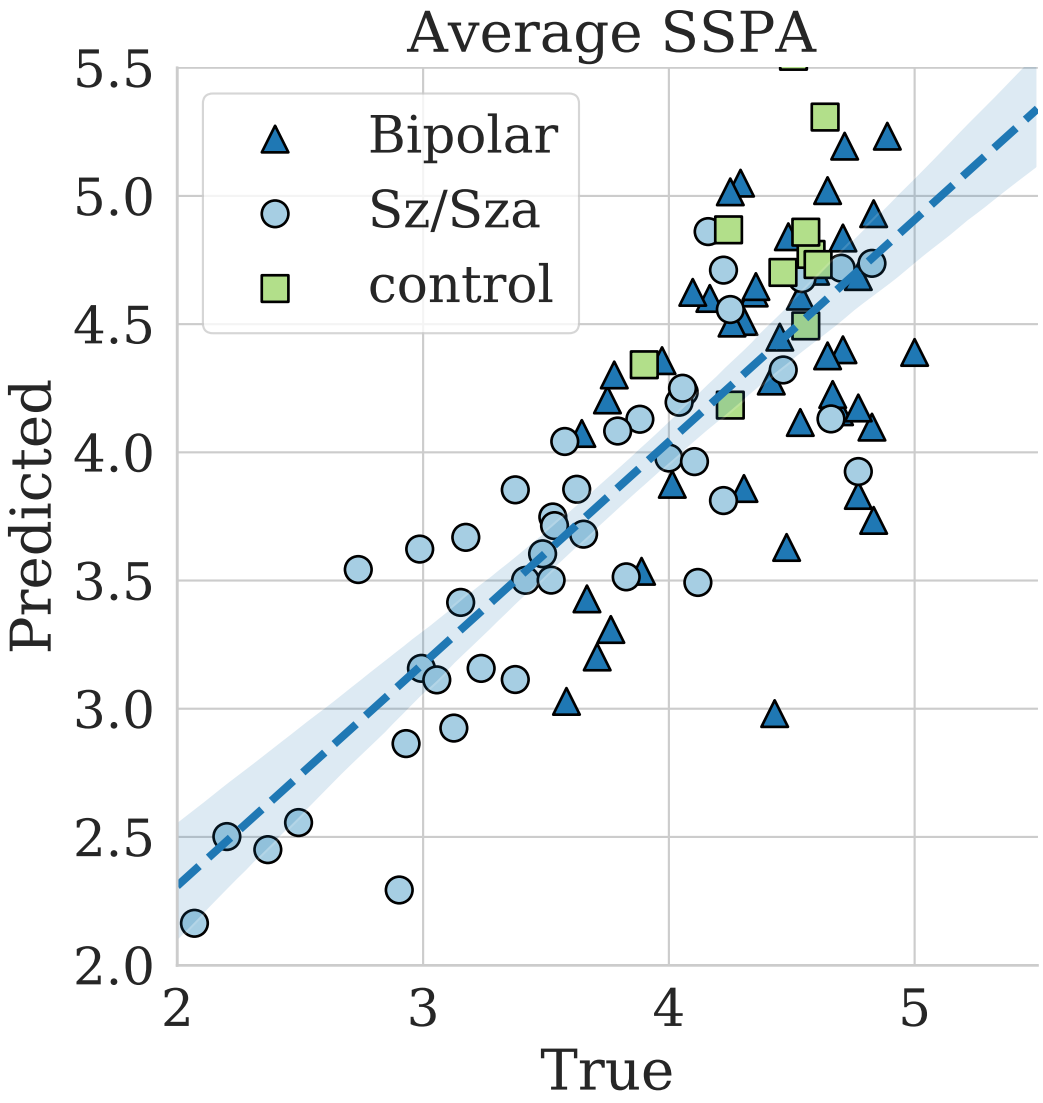


Figure 8: A linear regression model was fit using 25 out of the 73 semantic coherence and linguistic complexity features from the 109 subject responses to predict the SSPA scores. Correlation Coefficient = **0.752**, Mean Absolute Error = **0.330**, Root Mean Square (RMS) Error = **0.405**

with healthy controls. Finally, we considered parse tree statistics, computed using the Stanford Parser [115]. This includes the parse tree height and Yngve depth scores (mean, total, and maximum), a measure of embedded clause usage [61].

2.4.3 Results & Discussion

We first sought to determine a subset of language features (described in Section 2.4.2) from which we can accurately model the clinical SSPA scores. A total of 73 features were considered: 63 semantic features (7 statistical features $\times$ 3 sentence embedding types $\times$ 3 scenes) and 10 linguistic complexity features computed over all three scenes concatenated. Next, we aim to determine the predictive power of the selected subset of these features in separating the groups of interest (*i.e.* Sz/Sza, bipolar I disorder, and healthy control subjects). The regression and classification models built with these features were designed and tested using WEKA [116]. It is important to note that the SSPA itself is correlated to the clinical diagnosis and has been effective in differentiating groups of interest [23]. As a result, we note that using it to select features may result in overly-optimistic classification performance for the clinical vs. healthy control and Sz/Sza vs. bipolar disorder classification problems. However, due to the relative dearth of available data in this area, we performed this analysis on the same dataset.

2.4.3.1 Modeling SSPA Performance

We use a greedy stepwise search (with linear regression) through the feature space to determine the optimal subset of the features which accurately model the SSPA scores for all 109 subjects without considering the group variable. We down-selected to a set of 25 computed features out of the original 73. These are briefly summarized in Table 1,

Table 1: Selected features to model SSPA scores with a linear regression model, including ranking of overall importance for each feature. Italicized features were included in both the 25 feature and 15 feature classification problems.

Category	Features	Rank
Semantic Coherence	*BoW mean scene 3*	1
	INF minimum scene 3	2
	SIF 90th percentile scene 3	5
	INF maximum scene 2	7
	INF median scene 3	8
	BoW median scene 3	9
	BoW minimum scene 2	10
	BoW st. dev. scene 2	11
	BoW maximum scene 3	12
	INF st. dev. scene 3	13
	BoW maximum scene 2	18
	BoW 90$^{\text{th}}$ percentile scene 2	19
	BoW st. dev. scene 3	20
	BoW 90$^{\text{th}}$ percentile scene 3	21
	INF mean scene 3	22
	INF 10$^{\text{th}}$ percentile scene 3	23
	BoW 10$^{\text{th}}$ percentile scene 2	24
Lexical Diversity	*MATTR*	3
	Brunét's index	4
	Honoré's statistic	25
Lexical Density	*FUNC/W*	6
	UH/W	14
Syntactic Complexity	*Maximum Yngve depth*	15
	Mean length sent. (MLS)	16
	Parse tree height	17

and the resulting regression model (evaluated using leave-one-out) is shown in Figure 8. We notice that several of the coherence statistics for Scene 3 (negotiation with landlord) are particularly influential when tracking the assigned SSPA score with this model. Interestingly, the top three coherence statistics include a bag-of-words average of *word2vec* vectors (BoW mean scene 3), an *InferSent* sentence encoding (INF minimum scene 3), and a SIF embedding (SIF 90$^{\text{th}}$ percentile scene 3), indicating a variety of

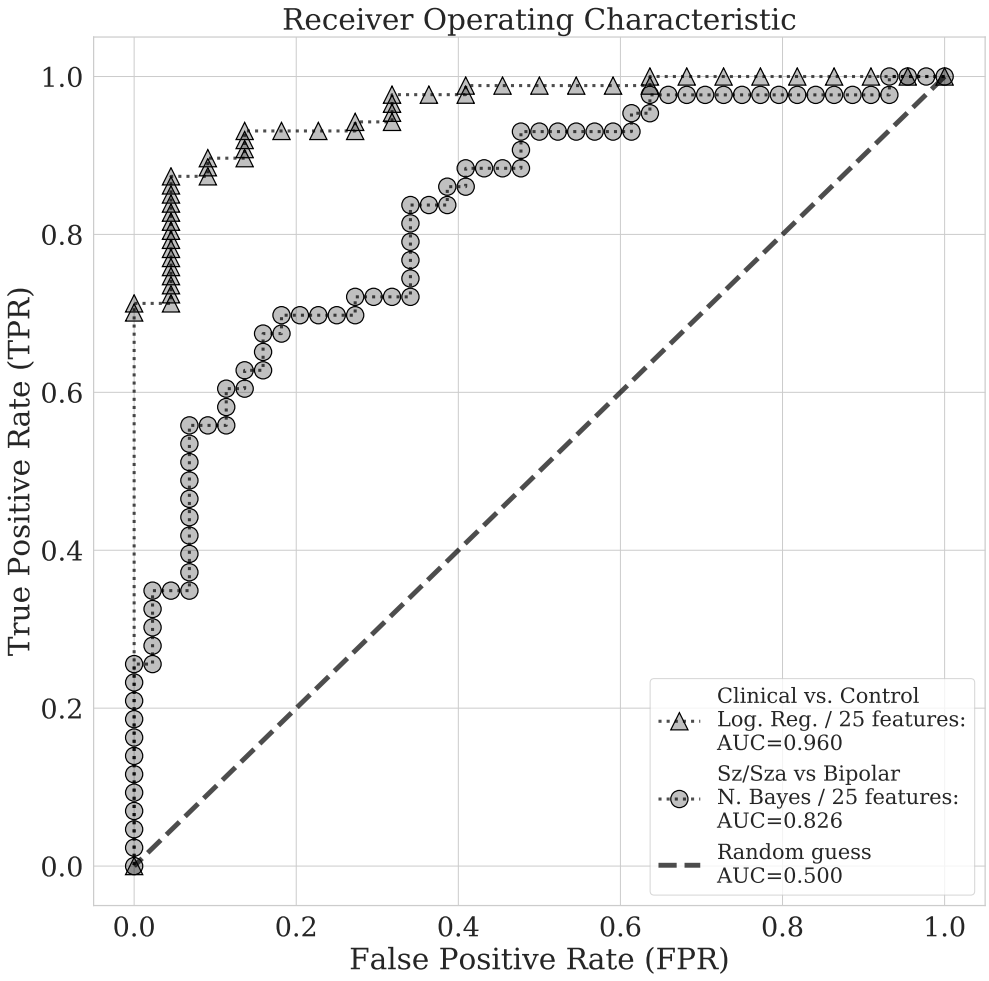


Figure 9: Selected receiver operating characteristic (ROC) curves for both binary classification tasks. For clinical vs control classification, TPR indicates correctly classifying a clinical subject and FPR indicates falsely classifying a control subject as clinical. For Sz/Sza vs bipolar classification, TPR is correctly classifying an Sz/Sza subject and FPR is falsely classifying a bipolar subject as Sz/Sza.

embeddings and range of statistics all provide useful information in predicting SSPA performance. We also note that a variety of lexical diversity (MATTR, Brunét's index), lexical density ($^{\text{FUNC}}$/w, $^{\text{UH}}$/w) and syntactic complexity (maximum Yngve depth)

Table 2: Confusion matrices for binary classification results with logistic regression (LR) and naïve Bayes(NB) classifiers with a 25 feature and 15 feature subset. a For clinical vs control classification, LR with 25 features works best at differentiating groups. b For Sz/Sza vs bipolar classification, LR using a 25 feature subset works poorly. NB provides more consistent results, even when the feature set is reduced.

(a) Clinical vs Control

Log. Reg.		True group: Clinical	Control
25 feat.	Clinical	78	3
	Control	9	19
		AUC = 0.960	
		Clinical	Control
15 feat.	Clinical	79	10
	Control	8	12
		AUC = 0.882	
N. Bayes		Clinical	Control
25 feat.	Clinical	73	2
	Control	14	20
		AUC = 0.908	
		Clinical	Control
15 feat.	Clinical	76	5
	Control	11	17
		AUC = 0.873	

(b) Sz/Sza vs Bip.

	True group: Sz/Sza	Bipolar
Sz/Sza	30	14
Bipolar	13	30
	AUC = 0.700	
	Sz/Sza	Bipolar
Sz/Sza	30	10
Bipolar	13	34
	AUC = 0.796	
	Sz/Sza	Bipolar
Sz/Sza	30	11
Bipolar	13	33
	AUC = 0.826	
	Sz/Sza	Bipolar
Sz/Sza	31	11
Bipolar	12	33
	AUC = 0.803	

measures are among the most influential, confirming the benefit of a complementary set of language measures.

2.4.3.2 Identification of Schizophrenia and Bipolar Disorder

Next, we aim to determine the ability of this subset of language features to correctly predict which subjects fall into the groups of interest. We performed two separate classification tasks: (1) separation of the clinical and healthy control groups, (2) separation within the clinical group between Sz/Sza subjects and bipolar I subjects.

Both a logistic regression (LR) and a naïve Bayes (NB) classifier were trained in each case using leave-one-out cross validation to determine model parameters and performance. Then, we further down-selected this set to a group of 15 features and re-evaluated the performance of both classifiers.

The confusion matrices for the clinical and control group classification task are shown in Table 2a. As we can see, LR with all 25 selected features works best, with the area under curve (AUC) in the ROC plot being 0.960 (see Figure 9). In this case, 78 of 87 (89.7%) clinical subjects and 19 of 22 (86.7%) healthy controls were correctly identified in our leave-one-out evaluation. We also see comparable performance for the NB and LR models when the feature set is reduced to only the top 15 features that model SSPA scores, though AUC is lower than both models with 25 features.

Next, we consider a classification problem within the group of 87 clinical subjects, of which 43 are diagnosed with Sz/Sza and 44 are diagnosed with bipolar I disorder. We use the same feature subsets and same binary classifier models as in the previous task, trained and evaluated using leave-one-out cross-validation. From the confusion matrices in Table 2b, we see that NB performs better than LR when either a 25 feature or 15 feature subset are used, with the best AUC = 0.826 for NB with 25 features. The ROC curve for a 25-feature NB classifier is shown in Figure 9. Interestingly, LR with 25 features had the lowest performance on this task (AUC = 0.700).

LR typically performs better than NB when more data is available for training [117]; however in clinical applications data set size is often limited. This makes sense with respect to our study, as the dataset used in the Sz/Sza vs. bipolar I classification problem is smaller than the dataset used in the clinical vs control group classification problem. In this case, the LR model is prone to overfitting, as is evident by the fact that performance improves when the feature dimension is reduced. As expected,

the classifier performance is considerably worse than the clinical and control group classification problems, as the language differences between schizophrenia and bipolar patients are more difficult to distinguish, even for experienced clinicians. Considering this fact, we still see reasonable performance with only computed language measures and no additional clinical assessment.

2.4.4 Concluding Remarks

This work demonstrates the potential of computational linguistics to aid neuropsychiatric practice in the clinic. We believe it is critically important to tie computational methods to established clinical practice in order to bridge the gap between the latest developments in NLP, which motivated our feature selection using SSPA. Still, there are many directions in which we can take future work. The sentence embedding and coherence metrics computed in this study are by no means an exhaustive list of potential methods, and it is likely a more optimal easily computable feature set exists to model SSPA performance and classify groups of interest. In particular, we are interested in finding a more concise group of clinically relevant language features with which we can perform this analysis. Additionally, we can look at more language metrics within each subject group to further subtype and cluster individuals within each group based on language metrics. These methods can also be applied to clinical assessments beyond the SSPA tasks and for a wider variety of psychiatric conditions. Lastly, we would like to examine how classification and modeling of clinical test scores changes when computed features are used in conjunction with other clinical tests to model task performance and classification of groups.

In the following chapters, we will demonstrate how we have advanced this work

using language features inspired by the speech production model we first introduced in Section 2.1.

Chapter 3

A NOVEL MEASUREMENT MODEL FOR CLINICAL SPEECH ANALYSIS

In this chapter, we introduce the framework for the language analysis used in our studies. In Figure 10, we see the first two domains of language production from the Levelt framework that we first introduced in Figure 3. One of the challenges with operationalizing any computational language assessment framework is reliable measurement of the latent domains of interest; in our case these are the individual items under the "Conceptualization" and "Formulation" umbrellas in Figure 10. These are likely multidimensional constructs that have yet to be operationally defined in the literature.

Briefly, our measurement model consists of three parts: extraction of a set of low-level features that have been used in previous work, mapping of these features to the individual Levelt stages, and denoising of these features using principal components analysis (PCA) [118, 119, 120]. The denoising step is critical as there is converging evidence that out-of-the-box speech and language features are high-dimensional, variable, prone to confounding, and exhibit poor test-retest reliability [121, 122]. Furthermore, machine learning models built on top of these features exhibit poor external validity [123]. The Levelt model serves as a theoretical guide for grouping the less reliable low-level features that aim to represent similar constructs into composites. In the section that follows, we provide an overview of the methods to compute these low-level features and their composites through the PCA approach.

The details here describe how the measurement model framework was used in our SSPA Study [22] for individuals with schizophrenia and bipolar disorder. However,

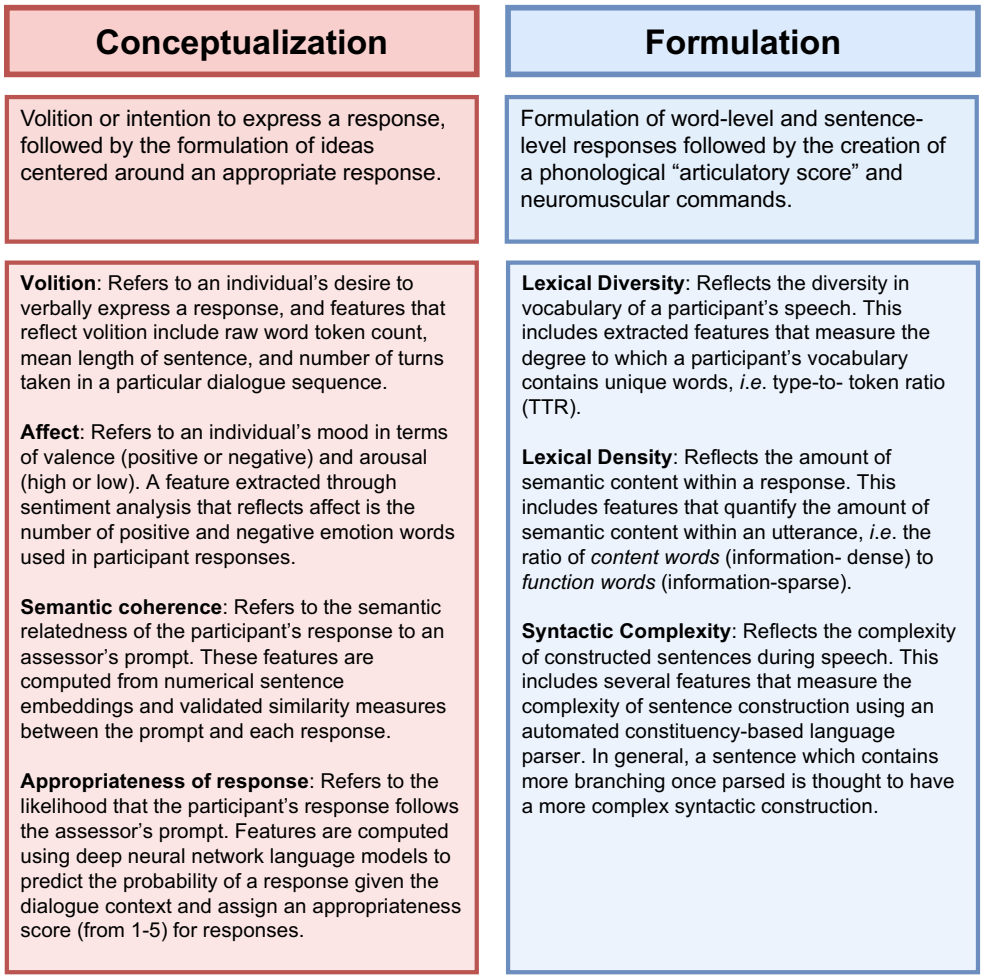


Figure 10: Two of the three stages of the speech production framework, a brief description of each stage (second row), and list of domains that characterize each stage (third row). We note that the The "articulation" stage is not included here because acoustic speech samples were not available for the transcripts studied (see Figure 3 for reference).

this framework can be generalized and adapted to studying speech and language analytics for any given neurological, cognitive, or mental disorder that impacts speech and language.

3.1 Framework for Spoken Language Production

In this work, we make use of a model proposed in [124] that characterizes spoken language production as a complex, multi-stage event consisting of three major stages:

1. *Conceptualization*: involves abstract idea formation and the intent or volition to communicate the idea.
2. *Formulation*: involves selection and sequencing of words and the precise linguistic construction of an utterance, along with a sensorimotor score for muscle activation.
3. *Articulation*: involves execution of this sensorimotor score by activation and coordination of speech production musculature (*i.e.* respiratory, phonatory, articulatory, *etc.*)

In [28] and [9], this framework was used to organize literature reviews of speech-based assessment of depression, suicidality, cognition, and thought disorders. In this work, we use it to define a measurement model for speech.

Our instantiation of the model herein is shown in Figure 10 and serves as a guide for a new representation of speech especially useful for clinical applications. It is important to note that the representation is directly tied to the speech elicitation task. In our work, we use the three scenes in the *Social Skills Performance Assessment* (SSPA) task, a role-play based assessment used to measure social competence skills. For this context, we can further divide the three stages into a set of more interpretable

domains, each of which can be measured by a constellation of lower-level features. Our goal is to identify a representation of language production that (1) is sensitive to impairment by thought and mood disorders (*i.e.* schizophrenia and BD) and (2) can be quantified and assessed by automated computational techniques in NLP. In this work we focus only on the first two stages, conceptualization and formulation, and leave the representation of the articulation domain for future work.

3.1.1 Conceptualization Stage Domains

The domains that fall under the conceptualization stage are described below. For each, we provide a high-level description and describe the low-level features that reflect that domain. These are then combined into a composite representation for each domain.

1. **Volition**: Refers to an individual's desire to verbally express a response, and features that reflect volition include raw word token count, mean length of response, and number of turns taken in a particular dialogue sequence.
2. **Affect**: Refers to an individual's mood in terms of valence (positive or negative) and arousal (high or low). An example feature extracted through sentiment analysis that reflects affect is the number of positive and negative emotion words used in participant responses.
3. **Semantic coherence**: Refers to the semantic relatedness of the participant's response to the assessor's prompt. These features are computed from numerical response embeddings and similarity measures between the prompt and each response.
4. **Appropriateness of response**: Refers to the likelihood that the participant's

response follows the assessor's prompt. Features are computed using deep neural network language models (*i.e.* BERT [73]) to predict the probability of a response given the dialogue context and assign an appropriateness score (from 1-5) for responses.

For each of these domains, the feature constellations are combined and reduced into a small set of composites using principal component analysis (PCA) [118]. This is done in an effort to combat the variability of the lower-level features, as explained in the manuscript.

3.1.2 Formulation Stage Domains

The domains that fall under the formulation stage are described below. For each, we provide a high-level description and describe the low-level features that reflect that domain. These are then combined into a composite representation for each domain.

1. **Lexical Diversity**: Reflects the diversity in vocabulary of a participant's speech. This includes extracted features that measure the degree to which a participant's vocabulary contains unique words, *i.e.* type-to-token ratio (TTR).
2. **Lexical Density**: Reflects the amount of semantic content within a response. This includes features that quantify the amount of semantic content within an utterance, *i.e.* the ratio of *content words* (information-dense) to *function words* (information-sparse)
3. **Syntactic Complexity**: Reflects the complexity of constructed sentences during speech. This includes several features that measure the complexity of sentence construction using an automated constituency-based language parser.

In general, a sentence which contains more branching once parsed is thought to have a more complex syntactic construction.

Except in the case of the *Affect* domain, all low-level features were computed by combining the transcripts from all three scenes in the SSPA using PCA, as described in the previous section. Since the emotional nature of Scenes 2 and 3 in the SSPA task were quite distinct, these features were computed independently for just scenes 2 and 3.

3.1.3 Schizophrenia and Bipolar Disorder in the Context of this Framework

Spoken language impairments in individuals with schizophrenia and bipolar disorder can be characterized in the context of the framework outlined in the previous section. [30]. Schizophrenia is a heterogeneous condition that is primarily associated with *formal thought disorder* (FTD), and can present with a variety of *positive* or *negative* symptoms [125]. Positive symptoms are those in which normal functions are expressed or distorted in excess and include hallucinations, delusions, and disorganized or incoherent "word salad" speech (*schizophasia*). We expect that positive symptoms associated with schizophasia will impact *semantic coherence* and *appropriateness of response* in objectively measurable ways. Negative symptoms refer to those that present some type of deficiency in individuals with schizophrenia, and may include lack of motivation (*avolition* or *amotivation*), apathy, flat affect, or negative thought disorder (*poverty of speech and language*). In terms of the framework in Figure 10, we expect these negative symptoms to have a measurable negative impact on *volition*, *affect*, *lexical density*, *lexical diversity*, and *syntactic complexity*. Individuals can also exhibit a subset of these symptoms at varying degrees of severity.

BD is characterized by the fluctuation between episodes of different *depressive* and *manic* mood states [30]. Each mood state is associated with a variety of symptoms that impact the speech and language output of that individual [126]. Manic episodes are characterized by *pressured speech*, which is described as excessively rapid and difficult to understand. It is also characterized by increased verbosity and *flight of ideas*, or quickly jumping from topic to topic in a disorganized manner [127]. Depressive mood states can result in exhibiting poverty of speech and language or increased pause times, similar to impairments associated with negative symptoms of schizophrenia. Therefore, within the defined framework, depressive speech will similarly primarily impact the conceptualization stage of language production, impacting our features tapping *volition* and *affect*. Manic speech can also impact the conceptualization stage, through excessive expression that may impact the *appropriateness of response* or *semantic coherence* in a given context; in the formulation stage, there may also be measurable impacts on *lexical density*, *lexical diversity*, and *syntactic complexity*.

In our work, we analyzed the transcripts of individuals with varying symptom severity for schizophrenia and bipolar disorder. As stated above, we aimed to identify language features that could both be associated with these particular impairments and could also be computed automatically with modern advancements in NLP.

3.2 Methods

Here, we provide a detailed overview of the computational methods used to extract linguistic features in the domains of interest within our framework. As previously stated, the focus in this study is on the linguistic *conceptualization* and *formulation* stages of language production, as an acoustic assessment of articulation is not possible

with purely textual transcript analysis. We leave this for future work. It is also important to note that each spoken utterance by participants in this SSPA study occurs in a conversational context, and the lower-level features used to modeling the conceptualization and formulation stages of language production consider this context.

3.2.1 Conceptualization Stage Measurement Model

As discussed in Section 3.1 and Figure 10, during the conceptualization stage, an individual forms an abstract idea of what he or she intends to speak. In a conversation, this can be measured in two ways: (1) by the total verbal output (which serves as a proxy for volition or motivation to speak), and (2) measures that objectify the appropriateness of a spoken response given the context.

3.2.1.1 Low-level Features for Volition

Volition is most simply measured by quantifying the verbal output of an individual, which in previous work has been shown to be predictive in other studies on schizophrenia, bipolar disorder, and Alzheimer's disease (AD) [17, 50]. In our work, for a given conversation, we used total words spoken (W), the number of participant turns (Turns/Dialogue), average number of words spoken in each turn (Tokens/Turn), the mean length of sentences (MLS), mean length of T-unit (MLT), and the mean length of clause (MLC) as a proxy for volition and motivation to speak.

3.2.1.2 Low-level Features for Affect

The Linguistic Inquiry and Word Count (LIWC) tool [42] is used for characterizing and categorizing the lexicon of a given body of text. The LIWC tool can classify words into categories related to affect, such as words associated with positive and negative emotions, which provides us with indirect measures of the sentiment of the speaker's language in a conversation. For our transcripts, the LIWC sentiment analysis was conducted to give absolute counts for words spoken by each participant in the following categories: {*negative emotions*, *positive emotions*, *death*, *sadness*, *anger*, *emotional ratio* (*positive* to *negative*)}. To simplify our analysis, the composite features computed for the Affect domain for scenes 2 and 3 were derived only from *negative emotions*, *positive emotions*, and the *emotional ratio* statistics for the transcripts of those scenes.

3.2.1.3 Low-level Features for Semantic Coherence

The deficiency in an individual's ability to form semantically coherent utterances is a hallmark of formal thought disorder associated with schizophrenia and BD. One way to quantify coherence is to study the semantic relationships between the dialogue context and each spoken utterance for a given participant. In NLP, semantics are computationally modeled with word or sentence *embeddings*, typically a high-dimensional vector representation of a body of text. Words or phrases used in similar semantic contexts are often represented closer together as measured by their *cosine similarity*, given in Equation (3.1),

$$\text{CosSim}(\mathbf{w}_1, \mathbf{w}_2) = \cos\theta = \frac{\mathbf{w}_1^{\mathrm{T}}\mathbf{w}_2}{\|\mathbf{w}_1\|_2 \, \|\mathbf{w}_2\|_2} \ , \tag{3.1}$$

where $\mathbf{w}_1$ and $\mathbf{w}_2$ are the vector representations of two bodies of text, θ represents the angle between the two embeddings, and $\|\cdot\|_2$ represents the Euclidean norm. Therefore, a cosine similarity can have a maximum value of 1 if the vectors are perfectly aligned, indicating identical semantic content. In this study, we are most interested in generating semantic vector representations of each utterance spoken by the assessor or participant in the SSPA task. As we did in our previous work [17], we considered unweighted averages of the *word2vec* [14], *smooth inverse frequency* (SIF) embeddings [74], and sentence representations generated by the *InferSent* sentence encoder [76].

Additionally, the recent language modeling technique, *Bidirectional Encoder Representations from Transformers* (BERT), proposed in [73], has improved computational performance across a variety of NLP tasks. BERT uses a transformer neural network architecture [128] to encode text with a large pre-trained language model that can be fine-tuned for increasing performance on particular tasks. Using a BERT implementation, we also followed the methodology in [129] to encode participant responses and the dialogue context to compute similarity scores.

The final reported features under this domain consist of a set of standard statistics (mean, median, maximum, minimum, standard deviation, 90^{th} percentile, and 10^{th} percentile) computed for each conversation using the similarity scores determined by each of the above methods.

3.2.1.4 Low-level Features for Appropriateness of Response

Similar to coherence, an inability to construct an appropriate response in a given context is an important feature in formal thought disorders. To quantify features that

can measure the degree to which a given response can be considered “appropriate" in a given dialogue context, we made use of BERT language modeling in two different ways, using the PyTorch [130] implementation of BERT from the *transformers* Python library from Huggingface [131]:

1. *Probability of response*: BERT is trained with a *next sentence prediction* task as one of its auxiliary objectives. For our purposes, we made use of the pre-trained BERT language model to compute the probability of each participant response given the previous utterance by the clinical assessor.
2. *Automated response scoring*: Here, we used the annotated, open-source, HUMOD (human movie dialogue) dataset [132]. The data set consists of of dialogue context-response pairs that contain both the actual responses from movie dialogue and randomly sampled responses for each context. Human annotators assigned each response a relevancy score from 1-5, resulting in a wide range of possible response scores for a given context. We fine-tuned the pre-trained BERT model by adding a regression layer on top of the pre-trained model to score each response for a given context, and then applied to the context-response pairs for each participant response in our transcripts to automatically assign a relevancy score from 1-5 to each response.

Again, for the response probabilities, or response scores described above, we computed a distribution of values for each conversation and summary of basic statistics for each feature was computed for each conversation (mean, median, maximum, minimum, standard deviation, 90^{th} percentile, and 10^{th} percentile of each distribution of values).

3.2.2 Formulation State Measurement Model

As discussed in Section 3.1, thought and mood disorders can also disrupt the formulation stage of language production, affecting an individual's choice of words and ability to form complex linguistic constructions. The computational methodologies we used to quantify the impact on language formulation fall into two large categories, those at the lexeme/word level (*i.e.* lexical diversity and density) and those at the sentence and utterance level (*i.e.* measures from parse trees constructed from the uttered sentences).

3.2.2.1 Low-level Features for Lexical Diversity

Lexical diversity is a measure of unique vocabulary usage. The simplest method by which this is quantified is the *type-to-token ratio* (TTR), previously defined in Equation (2.1). This is simply the ratio of unique words (*types*, V) to total words spoken (*tokens*, N). However, TTR tends to plateau for longer utterances and alternative methods exist to account for this length dependence. In addition to TTR, we also consider the following measures of lexical diversity which limit the length dependence:

- *Moving-average type-to-token ratio* (MATTR) [43]: a measurement of TTR that uses a sliding window of fixed length for a given body of text, averaged over the length of the text.
- *Brunét's Index* (BI) [44]: previously defined in Equation (2.2) in which the exponential reduces the dependence on the total length N. Lower values for BI indicate increased diversity.

- *Honoré's Statistic* (HS) [45]: previously defined in Equation (2.3), which emphasizes the use of words that are only spoken once (V_1)

3.2.2.2 Low-level Features for Lexical Density

Lexical density is a measure of the amount of information that is packaged within an utterance. This can be quantified by the content density, or ratio of information-dense *content words* (*i.e.* nouns, verbs, adjectives, adverbs) to information-sparse *function words* (*i.e.* prepositions, conjunctions, interjections, *etc.*). In our work, we used the Stanford part-of-speech tagger [48] to identify the content and function words. We used two measures as an inverse of lexical density, FUNC/w (ratio of function words to total words) and UH/w (ratio of interjections to total words), as in [17].

3.2.2.3 Low-level Features for Syntactic Complexity

To measure the degree of syntactic complexity in the language of each participant, we first concatenated all participant utterances and ignored the speech of the clinical assessor. Then, for each utterance, we used constituency-based parsing using the Stanford parser tool [115]. This allows automatic deconstruction of an utterance into its syntactic structure.

For constituency-based parsing, we use the Stanford parser [115] to decompose each sentence spoken by the participant. Then, Yngve scoring [61] is done for each sentence. An example of constituency-based parsing and Yngve scoring was shown previously for a simple sentence in Figure 4a. We considered parse tree statistics to

represent this domain. These include the parse tree height and Yngve depth scores (mean, total, and maximum), a measure of embedded clause usage [61].

3.2.3 Feature Composites via Principal Component Analysis

For each of the seven domains, we applied *principal component analysis* (PCA) by domain [118] to denoise the more variable low-level features. We began with the raw set of 43 computed features described in the previous section, organized by the domains which they represent. The number of principal components (PCs) used to represent each domain was selected such that they account for at least 85% of the variance of all features within that domain. As a result, we obtained 2 PCs for volition, 4 PCs for affect, 2 PCs for lexical diversity, 2 PCs for lexical density, 1 PC for syntactic complexity, 6 PCs for semantic similarity, and 4 PCs for appropriateness of response (a total of 21 features). Since several of the computed features co-vary with the raw number of words spoken, the PCA representation of the feature domains were provided to the model designer along with the raw count of word tokens (W) spoken by the participant in each dialogue to use for model development. This feature set was used to develop several prediction models.

Chapter 4

UPSTREAM AND DOWNSTREAM EVALUATION OF MENTAL HEALTH AND FUNCTIONAL COMPETENCY

In this chapter, we make use of the measurement model framework introduced in Chapter 3 and apply it to performing the relevant analyses for our study on schizophrenia and bipolar disorder with the SSPA transcripts. Importantly, in accordance with Aim (2) of this dissertation, we expand upon the vast majority of previous literature by performing a holistic study that includes an analysis and assessment of both upstream and downstream outcomes from using language output as the measurement medium (See Figure 2). *Upstream* analyses include those that relate to mental health status, *i.e.* diagnosistic classification, symptom severity ratings, and cognitive assessments. *Downstream* analyses include measures of social and functional competency.

While most previous research has primarily focused on the upstream classification problem using speech and language features, we propose that this full set of analyses is far more useful in a clinical setting, in which the goal of medical care and intervention is to improve overall quality of life for afflicted individuals.

4.1 Data Used for Model Development and Model Evaluation

For all models developed here, we used the same sample of participant interviews. Data from a total of 281 participants with a clinical diagnosis of either schizophrenia/schizoaffective (Sz/Sza) disorder ($n = 140$) or bipolar disorder (BD) ($n = 141$) and 22 healthy controls were used in this study. Every participant was subject to

Table 3: Participant demographics for the training set (used during cross-validation) and the out-of-sample test set (for model evaluation).

Training		Sample Size (Gender)	Age	Years of Education
	Sz/Sza	98 (37 F, 61 M)	μ: 51.27 σ^2: 10.10 R: 25-75	μ: 14.43 σ^2: 2.65 R: 6-20
	BD	98 (51 F, 47 M)	μ: 47.45 σ^2: 13.23 R: 18-80	μ: 16.08 σ^2: 2.20 R: 11-20
	Control	11 (3 F, 7 M, 1 undisclosed)	μ: 38.40 σ^2: 10.42 R: 23-52	μ: 16.40 σ^2: 1.96 R: 13-18
Out-of-Sample		Sample Size (Gender)	Age	Years of Education
	Sz/Sza	43 (18 F, 24 M, 1 undisclosed)	μ: 50.26 σ^2: 10.83 R: 23-78	μ: 13.73 σ^2: 2.76 R: 8-18
	BD	42 (18 F, 24 M)	μ: 50.57 σ^2: 11.83 R: 21-75	μ: 16.29 σ^2: 1.78 R: 12-20
	Control	11 (8 F, 3 undisclosed)	μ: 43.63 σ^2: 10.90 R: 24-57	μ: 16.75 σ^2: 1.49 R: 14-18

extensive clinical evaluations that consisted of neurocognitive batteries, symptom ratings, social, and functional assessments. Participant demographics (*i.e.* gender, age, years of education) are summarized in Table 3. A summary of the statistics for which participants were evaluated are shown in Table 4 and Table 7.

4.1.1 Language Samples:

Language samples from each participant were elicited via the *Social Skills Performance Assessment* (SSPA) task [21]. The SSPA is a simple-to-administer role-playing test that can serve as a measurement of skills related to social competence. Participants in the study are asked to act out the following three "scenes" with a clinical assessor:

- *Scene* 1 (practice): the participant plans a weekend activity with a friend ($\sim$ 1 minute)
- *Scene* 2 (scored): the participant introduces a new neighbor to his or her neighborhood ($\sim$ 3 minutes)
- *Scene* 3 (scored): the participant negotiates with a difficult landlord to fix a leak in his or her apartment ($\sim$ 3 minutes)

Scenes 2 and 3 are individually scored on a scale from 1 to 5 across a variety of dimensions, such as overall interest/disinterest, affect, negotiation ability, fluency, *etc.*An overall score for each scene can be computed by averaging the scores across each dimension of interest for each scene.

The SSPA was administered by trained researchers at the Psychology department at Queen's University at Kingston, Ontario, Canada. The samples were manually transcribed.

4.1.2 Development / Test Split:

The data was randomly split into two sets, a development set and a test set. A table of descriptive statistics for all relevant outcome measures is shown in Table 4. The development set was used by the algorithm developer to create the models. Importantly, the algorithm developer did not have access to the test set at any point during model development. Once the model was fixed, it was shared with us, and we evaluated the performance of the model on the test set. The table also includes a summary of the training and test set samples along with relevant statistics.

4.1.3 Data Analysis

Two researchers (Author 1 and Author 5 from [22]) participated in the training and testing of a series of prediction models. Author 1 split the data into a training and test set. Only the training set was provided to Author 5, who trained the final models. These models were then fixed and shared with Author 1, who evaluated their performance on the held-out test set.

Several regression models were developed to predict the upstream and downstream clinical scales of interest, including the average SSPA score, SLOF scores (functional composite and the activities, interpersonal relations, and work skill subscales), neurocognitive composite, and PANSS positive and negative symptom averages. Additionally, two classification models were developed in order to predict if individuals belonged to different diagnostic groups. The first attempted to classify healthy controls against those that were part of a clinical diagnostic group (Sz/Sza or BD); the next attempted to correctly classify those in clinical groups into either Sz/Sza or BD.

4.1.4 Model Training

4.1.4.1 Linear Regression Prediction Models

Author 5 followed the following process for developing each of the upstream and downstream regression models. For each outcome, the goal was to create a model that was as simple as possible to avoid overfitting and included principal components instead of individual metrics to avoid unnecessary complexity, low-level feature variability, and collinearity among predictors.

Before the model-fitting process began, it was observed that the single feature W (total number of word tokens) was correlated with several of the other features and dependent variables.

Therefore, W or a square root transformation ($\sqrt{W}$) was included in all models. Linear regression models were fit starting with W or $\sqrt{W}$, and each principal component was added to the model one-at-a-time. If the prediction accuracy increased, the component was kept; if the prediction accuracy remained the same or decreased, the component was removed. Model accuracy was measured using the mean absolute error. For each new component that was added, the predicted and observed outcome scores were plotted to detect any non-linearities or outliers. If the inclusion of a predictor resulted in a non-linear prediction or outliers, a variable transformation was attempted, such as a logarithm or square root. Therefore, several competing models were considered for each outcome. The final models were selected based on: (1) minimizing the mean absolute error, (2) maximizing the correlation between the predicted and observed scores, (3) maintaining the smallest number of predictors in the model as possible, including W and principal components, and (4) no outliers.

This was all performed using leave-one-out cross-validation on the training set only. The full results are seen later for the training (cross-validation) and out-of-sample test sets in Table 5 and Table 8.

4.1.4.2 Diagnostic Classification Prediction Models

For the diagnostic classification models, two models were built: one that predicted clinical vs. healthy control and one that predicted BD vs. Sz/Sza. Logistic regression was used to predict the binary outcome, and the predictors included W and a subset of the principal components. The predicted score was the logit. Model performance was evaluated using the receiver operating characteristic area under the curve (ROC AUC) using leave-one-out cross-validation on the training data. The same process as described above was used, where each principal component was added one-at-a-time. Finally, once the final model was fixed, to assign a predicted class to each participant, a threshold for the logit score was set based on the highest unweighted average recall score. The results in the accompanying Table 6 show the ROC AUC for the two models on the out-of-sample data along with weighted precision and recall, and $F1$ score.

4.2 Assessment of Mental Health Status - The *Upstream* Problem

4.2.1 Final Upstream Models

For reference, we list the final models that were developed from the training samples provided to a third-party biostatistician (for unbiased model development).

Classification Model 1 (Clinical vs. Control):

$$\begin{aligned}\text{logit} = 15.43028 - 0.01199 * W - 0.92726 * \ &\text{Appropriateness.PC1} \\ - 1.11421 * \text{Semantics.PC4} + 1.74472 * &\text{Semantics.PC3}\end{aligned}$$

$$\text{class} = \begin{cases} \text{clinical} & \text{logit} \leq 0.48673509 \\ \text{control} & \text{logit} > 0.48673509 \end{cases}$$

Classification Model 2 (Sz/Sza vs. BD):

$$\text{logit} = 3.732518 - 0.005106 * W + 0.465529 * \text{Semantics.PC4}$$

$$\text{class} = \begin{cases} \text{BD} & \text{logit} \leq 0.06590161 \\ \text{Sz/Sza} & \text{logit} > 0.06590161 \end{cases}$$

Neurocognitive Composite Score Prediction:

$$\begin{aligned}\text{Cog Comp} = &- 4.1254 + 0.1280 * \sqrt{W} - 0.1539 * \text{Lex.Div.PC2} \\ &+ 0.1607 * \text{Affect.PC2}\end{aligned}$$

Table 4: Participant statistics for clinical upstream assessments of neurocognition and symptoms. Healthy control participants were not evaluated and are excluded.

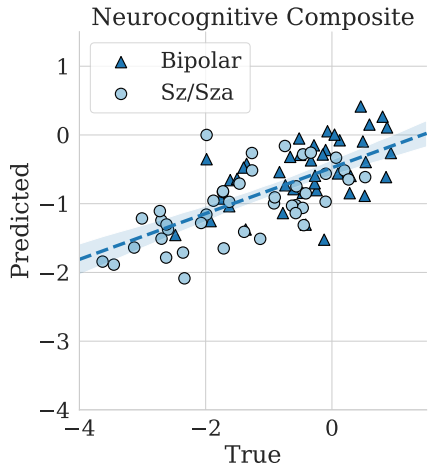
		Sz/Sza		**BD**	
		Model training	Out-of-sample	Model training	Out-of-sample
Neurocog. Composite	n	97	43	98	42
	μ	-1.13	-1.43	-0.34	-0.42
	σ^2	1.00	1.10	0.85	0.87
	R	(-3.27)-$(+0.76)$	(-3.63)-$(+0.53)$	(-2.40)-$(+1.41)$	(-2.48)-$(+0.93)$
PANSS					
Pos. Symptoms Mean	n	98	43	98	42
	μ	2.27	2.44	1.53	1.40
	σ^2	0.87	0.76	0.59	0.47
	R	1.00-4.86	1.14-4.00	1.00-3.29	1.00-3.14
Neg. Symptoms Mean	n	98	43	98	42
	μ	2.38	2.41	1.26	1.39
	σ^2	1.12	1.22	0.37	0.48
	R	1.00-5.86	1.00-6.14	1.00-2.43	1.00-3.00

PANSS Positive Symptoms Mean Prediction:

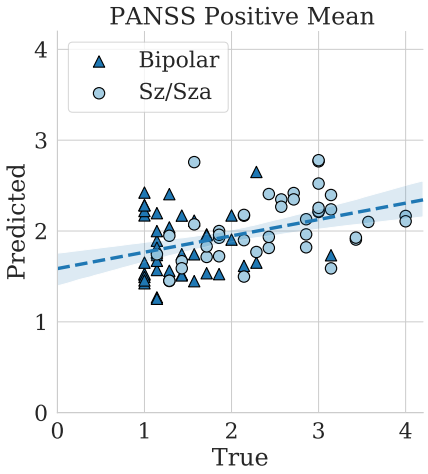
$$\begin{aligned}\text{PosMean} =& 3.10246 - 0.04630 * \sqrt{W} - 0.14257 * \text{Appropriateness.PC1} \\ &+ 0.07573 * \text{Lex.Div.PC2} - 0.17325 * \text{Volition.PC2} \\ &+ 0.10953 * \text{Appropriateness.PC2}\end{aligned}$$

PANSS Negative Symptoms Mean Prediction:

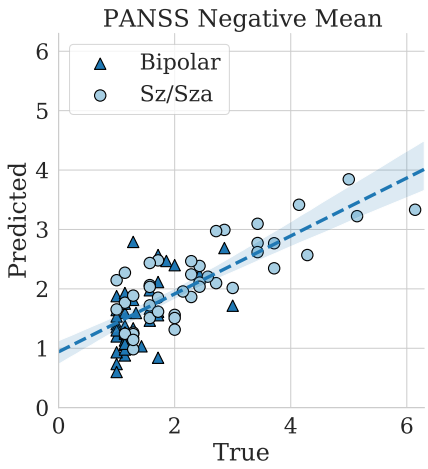
$$\text{NegMean} = 5.7816 - 0.1499 * \sqrt{W}$$



(a) Neurocognitive composite score



(b) PANSS positive symptoms mean



(c) PANSS negative symptoms mean

Figure 11: Demonstrated fit of linear regression models on out-of-sample transcripts for predicting upstream neurocognition and symptom rating measurements. Only includes Sz/Sza and BD groups as symptom ratings and neurocognition variables were not assessed for healthy controls. The associated regression model performance statistics are available in Table 5.

Table 5: This table shows the performance of the linear regression models developed to predict performance on *upstream* outcomes in neurocognition and symptom assessment. The table shows the performance of the models in terms of coefficient of determination (R^2), the Pearson correlation coefficient (PCC), mean absolute error (MAE), and mean squared error (MSE) between the predicted and actual outcomes for each task for both the samples used for model development (cross-validation) and new unseen transcripts (out-of-sample).

		R^2	PCC	MAE	MSE
Neurocog. Composite	Cross-validation	0.386	0.621	0.623	0.618
	Out-of-sample	0.442	0.674	0.682	0.668
PANSS:					
Pos. Symptoms Mean	Cross-validation	0.247	0.497	0.515	0.580
	Out-of-sample	0.258	0.509	0.492	0.559
Neg. Symptoms Mean	Cross-validation	0.516	0.718	0.487	0.509
	Out-of-sample	0.570	0.767	0.476	0.501

4.2.2 Experiments & Results

Our upstream analysis consists of three different models. We use the language feature domain representations to predict the neurocognitive composite score (a combination of several cognitive assessments) and the average values of the positive and negative symptom scale ratings (PANSS positive and negative averages). Additionally, we also use our reduced feature sets to classify between the clinical group and healthy Controls and to classify the clinical group into the corresponding diagnostic groups (*Sz/Sza*, *BD*).

4.2.2.1 Neurocognitive Composite Score Prediction

Schizophrenia, schizoaffective disorder, and bipolar disorder are known to negatively impact neurocognition to varying degrees for afflicted individuals. A composite neurocognitive score was computed and reported for each clinical participant (excluding healthy control participants) using the methods previously described in [23]. In summary, the composite score consists of eight well-known neurocognitive batteries: the Rey Auditory Verbal Learning Test [133], Trail Making Test [134], letter-number span test from the Weschler Adult Intelligence Scale (WAIS) [135], Wisconsin Card Sorting Test [136], digit-symbol coding test from the WAIS, a semantic fluency test [137], d' from the Continuous Performance Test-Identical Pairs Version, 4-digit condition [138], and the reading subtest of the Wide-Range Achievement Test, 3$^{\text{rd}}$ edition [139]. Standardized z-scores from these tests were used to compute a composite score.

The distribution of neurocognitive composite scores are summarized in Table 4. The performance of the regression models used to predict the composite neurocognitive score is shown in Table 5.

4.2.2.2 Positive and Negative Symptoms Scale (PANSS) Rating Predictions

The Positive and Negative Symptoms Scale (PANSS) assessment [140] consists of seven items that measure the severity of positive symptoms and seven items that measure the severity of negative symptoms. We use the average value of the positive symptom values and negative symptom values for each participant in accordance with the model proposed in [141]. The distribution across our participant groups for these

Positive Symptoms Mean and *Negative Symptoms Mean* values are shown in Table 4. Once again, healthy controls are excluded.

The performance of the linear regression predictive model on the training sample and out-of-sample participants is summarized as well in Table 5.

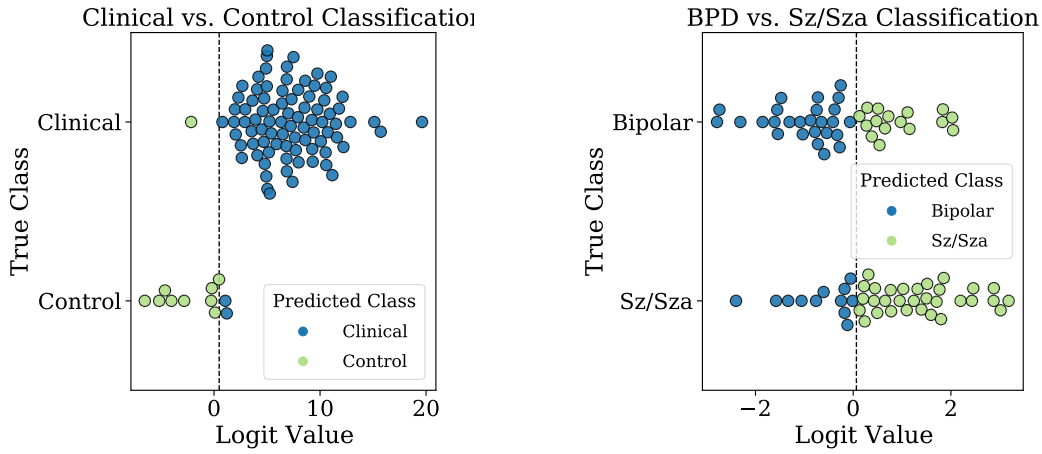
4.2.2.3 Diagnostic Group Class Prediction

The next set of experiments in our upstream analysis aimed to correctly identify the diagnostic group to which each participant belongs using the computed language features. This was accomplished by fitting logistic regression binary classifiers on the training set of participant transcripts in two separate experiments.

The first task (*Clinical vs. Control*) was to identify if participants fall into a clinical diagnostic group (Sz/Sza or BD) or are a healthy control. Since there is a large discrepancy in the overall number of clinical participant transcripts ($n = 195$ in the training data set) when compared to healthy controls ($n = 11$ in the training data set), we employed the SMOTE [142] data augmentation technique to generate synthetic control samples and to over-sample the minority class during cross-validation to counter the imbalance.

The next task (*BD vs Sz/Sza*) aimed to fit a similar logistic regression binary classifier to differentiate between the individuals that belonged to each of the two groups within the clinical transcript samples. Since the Sz/Sza and BD classes are quite balanced in support, no data augmentation or over-sampling was used.

The results from both classification experiments are summarized for the cross-validation and out-of-sample participants in Table 6 and Figure 12. The results are shown with weighted averages for precision, recall, and F1 score for correctly



(a) (b)

Figure 12: Out-of-sample logistic regression classification results for the two models that were developed: (a) Clinical Participants vs. Healthy Controls, and (b) BD Participants vs. Sz/Sza Participants

predicting each class (weighted by the support of that class). We also show the area-under-curve (AUC) for the receiver operating characteristic (ROC) curve to evaluate the performance of the classifier.

4.2.3 Discussion

Upstream assessment included the prediction of neurocognition measures (through the neurocognitive composite score), prediction of PANSS negative and positive symptom severity, and classification of participants into their known diagnostic groups (Sz/Sza, BD, or Healthy Control). Most previous work in this field has focused primarily on the diagnostic classification task. However, we aimed here to additionally study the ability of language to serve as a predictor for neurocognition and a measure of symptom severity; the hope is that if clinicians are provided a more complete

Table 6: This table shows the results from the two upstream logistic regression classification experiments performed on the language samples collected from the SSPA task. The first aims to differentiate between *clinical* (Sz/Sza or BD) participants and healthy *control* participants, whereas the second aims to differentiate between the Sz/Sza and BD participants. The results are reported with the confusion matrix, receiver operating characteristic area-under-curve (AUC), and a weighted average of precision, recall, and F1 score for each class prediction. Results are provided for both the cross-validation and out-of-sample participants for both experiments.

Clinical vs Control

Cross validation

	Clinical Predicted	*Control Predicted*
Clinical True	193	3
Control True	3	8

Precision = 0.971, Recall = 0.971
F1 = 0.971, AUC = 0.856

Out-of-sample

	Clinical Predicted	*Control Predicted*
Clinical True	84	1
Control True	2	9

Precision = 0.968, Recall = 0.969
F1 = 0.968, AUC = 0.903

BD vs Sz/Sza

Cross validation

	BD Predicted	*Sz/Sza Predicted*
BD True	74	24
Sz/Sza True	29	69

Precision = 0.730, Recall = 0.730
F1 = 0.729, AUC = 0.730

Out-of-sample

	BD Predicted	*Sz/Sza Predicted*
BD True	26	16
Sz/Sza True	12	31

Precision = 0.672, Recall = 0.671
F1 = 0.670, AUC = 0.670

picture of the underlying upstream conditions, they can use this information to guide better, targeted treatment regimens.

4.2.3.1 Neurocognitive Composite

Using only the PCA representation of the computed language features, we were able to show reliably consistent performance in predicting the neurocognitive composite score for both the cross-validation and out-of-sample participants. The most significant features for predicting this neurocognition variable are $\sqrt{W}$, *lexical diversity*, and *affect*.

While there is significant correlation between the predicted neurocognitive score

and the true score in both sets, the relative error of the predictions was found to be quite high. This suggests that there is significant potential to assess neurocognitive variables through automated language analysis, but we may be able to do more granular predictions for specific scales of neurocognition more accurately with targeted and varied language samples.

4.2.3.2 Positive and Negative Symptoms Scale (PANSS)

Next, we look at the ability of the PCA representation of the language features to predict symptom severity as defined by the positive and negative symptoms scale (PANSS). We took averages of the positive symptom ratings and negative symptom ratings and attempted to fit a model to predict each individually.

From the results in Table 5, it is apparent that the language features computed from the SSPA transcripts best lend themselves to predict the severity of negative symptom severity. The prediction model for the PANSS negative symptoms mean only used $\sqrt{W}$ as a feature, meaning that negative symptom severity is highly dependent on volition. Since negative symptoms are most associated with factors like poverty of speech, this tracks with what we know about the language of schizophrenia and bipolar disorder with severe negative symptoms.

Positive symptoms proved slightly harder to predict from our transcripts, with lower correlation between predictions and actual ratings and slightly higher prediction errors. The feature domains used in the prediction model were *volition*, *appropriateness of response*, and *lexical diversity*. As positive symptoms are more varied and associated with factors like flight of ideas, tangentiality, and many others, it is reasonable that the prediction model would be more complex and would consider more relevant feature

domains. It is also reasonable that the predictive performance itself is limited by the types of language samples that are collected with the SSPA task. Again, it is likely that different types of language samples (such as those from a free-form narrative story-telling task) would provide better opportunity to compute features that are more associated with positive symptoms scale ratings.

4.2.3.3 Classification Results

As shown in previous work [17], language features that are computed from the SSPA task transcripts have demonstrated predictive value in correctly identifying the diagnostic group to which a participant belongs. Like the work in [17], we performed the same two experiments (*Clinical vs. Control* and *BD vs. Sz/Sza*) with some key modifications. First, we used a much larger sample of participants in the current study and set aside a significant number (on which the model was never trained) to evaluate out of-sample performance. Second, we simplified and drastically reduced the number of features to our interpretable PCA representation motivated by our speech and language production model.

The results from these experiments were summarized in Table 6. As we can see, both models perform very similarly for the cross-validation of the training transcripts and evaluation on the out-of-sample transcripts. Additionally, in all cases, all classification metrics (AUC and weighted averages of precision, recall, and F1 score) show similar performance for correctly determining the class of the participant.

As previously shown in [17], the first classification experiment (*Clinical vs. Control*) demonstrates a much stronger ability for correct classification with our simple model. This is expected, since we typically see the largest differences in performance on the

SSPA task between healthy individuals and those with either type of impairment. Schizophrenia, schizoaffective disorder, and bipolar disorder can be thought to exist on a spectrum of decreasing severity for the prevalence of symptoms; therefore, it is expected that incorrect classification occurs mostly between those who are borderline between BD and healthy control for the first experiment. We can take a closer look at the confusion matrices in Table 6 for the *Clinical vs. Control* model. There are three incorrectly classified clinical participants (*i.e.* falsely predicted healthy controls) in the training sample, two of whom are in the BD group. There is also one incorrectly classified clinical participant from the test sample, who also belongs to the BD group. This lends support to our hypothesis that confusion is most likely to occur between milder forms of BD and healthy controls for this experiment. The model considered W, *appropriateness of response*, and *semantics* as the most important features for differentiating between healthy individuals and those with an impairment, which tracks well with what we know about the impact that Sz/Sza and BD can have on language output.

The differentiation between the clinical diagnostic groups is a challenging problem for clinicians, as Sz/Sza and BD are often confused for each other and misdiagnosed [143]. Therefore, it is reasonable that our model shows decreased performance and an increased number incorrectly classified individuals between these two clinical groups. Again, we see similar performance for correctly classifying the out-of-sample participants using the models developed with the cross-validation data. Additionally, we see consistently good performance for correct classification of each condition individually. This model considered W and *semantics* as the important differentiating features between the Sz/Sza and BD groups; this suggests that the largest differences in language between those with Sz/Sza and BD are related to spoken language output

and semantics (*i.e.* poverty of speech and semantically irrelevant speech are most common in schizophrenia).

4.3 Assessment of Social and Functional Competency (The *Downstream* Problem)

4.3.1 Data Description

For all downstream model development, the data used for model development and model evaluation were the same as previously described in Section 4.1. A summary table of the distribution of groups of interest for downstream analysis is shown in Table 7.

4.3.2 Final Downstream Models

The final trained downstream models are listed here for reference.

Average SSPA Score:

$$\begin{aligned}\text{SSPA Avg. Score} = {} & .15 * \sqrt{W} - 0.089777 * \text{Lex.Div.PC2} \\ & + 0.085640 * \text{Appropriateness.PC1} - 0.115232 * \text{Volition.PC1} \\ & - 0.082777 * \text{Affect.PC2} - 0.086493 * \text{Appropriateness.PC4} \\ & - 0.054675 * \text{Appropriateness.PC3}\end{aligned}$$

SLOF - Fx Composite

$$\text{SLOF Fx} = 5.9922 + 0.2582 * \sqrt{W} + 0.1645 * \text{Semantics.PC2}$$

SLOF - Activities Subscale

$$\text{SLOF Activities} = 2.82305 + 0.06852 * \sqrt{W} + 0.03072 * \text{Appropriateness.PC1}$$

SLOF - Interpersonal Skills Subscale

$$\text{SLOF Interpersonal} = 2.09122 + 0.08016 * \sqrt{W} + 0.10498 * \text{Semantics.PC2}$$

SLOF - Work Skills Subscale

$$\text{SLOF Work} = 1.2083 + 0.1046 * \sqrt{W} - 0.1129 * \text{Semantics.PC4}$$

4.3.3 Experiments & Results

For all predictive analyses, the regression models were developed and optimizing using leave-one-out cross-validation (LOOCV) on only the training samples; the best performing model was selected and fixed and then subsequently evaluated on the test samples. All models were built using the features described in Section 3.2. We controlled for total number of words spoken in each model, as several of the features co-varied with it.

Note that healthy control target scores are only available for the SSPA prediction model. For all other analyses we are only considering the Sz/Sza and BD samples.

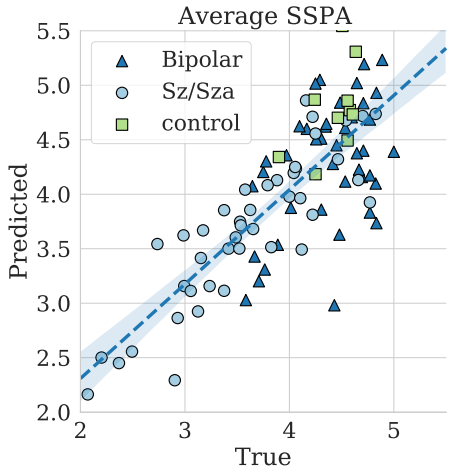
Table 7: Participant statistics for clinical *downstream* assessments of social and functional competency. Note that healthy control participants were only evaluated on the SSPA task.

		Sz/Sza		**BD**		**Control**	
		Model training	Out-of-sample	Model training	Out-of-sample	Model training	Out-of-sample
SSPA Avg.	n	97	43	98	42	11	11
	μ	3.79	3.61	4.42	4.37	4.48	4.47
	σ^2	0.73	0.70	0.39	0.40	0.24	0.26
	R	1.11-5.00	2.07-4.83	3.58-5.00	3.58-5.00	4.14-4.88	3.90-4.88
SLOF:							
Interpersonal	n	98	42	97	42	-	-
	μ	3.93	3.95	4.49	4.44	-	-
	σ^2	0.85	0.94	0.67	0.66	-	-
	R	1.57-5.00	1.29-5.00	2.14-5.00	2.57-5.00	-	-
Activities	n	96	42	98	42	-	-
	μ	4.44	4.31	4.82	4.82	-	-
	σ^2	0.64	0.64	0.28	0.30	-	-
	R	1.73-5.00	2.55-5.00	3.45-5.00	3.50-5.00	-	-
Work	n	93	41	98	41	-	-
	μ	3.56	3.34	4.37	4.26	-	-
	σ^2	1.03	0.87	0.84	0.82	-	-
	R	1.40-5.00	1.83-5.00	1.67-5.00	2.33-5.00	-	-
Fx Composite	n	92	41	97	41	-	-
	μ	11.94	11.57	13.67	13.50	-	-
	σ^2	2.04	1.98	1.55	1.38	-	-
	R	5.68-15.00	6.00-15.00	8.55-15.00	10.32-15.00	-	-

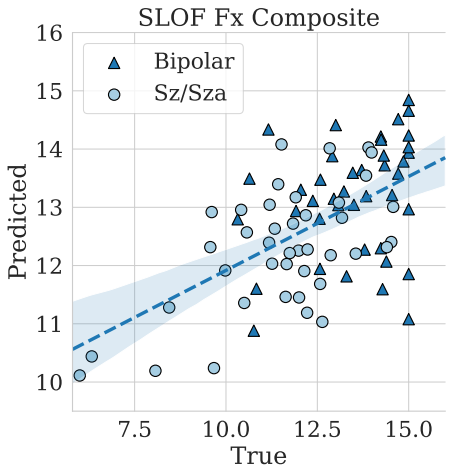
4.3.3.1 Social Skills Performance Assessment Score Prediction

We develop and evaluate regression models for predicting SSPA scores [21] using the reduced representation of our feature domains. Descriptive statistics for SSPA scores are shown in Table 7.

A summary of the cross-validation (for model development) and out-of-sample test performance for SLOF prediction models is shown in the bottom half of Table 8.



(a) SSPA average score between scenes 2 and 3.



(b) SLOF Fx composite score

Figure 13: Demonstrated fit of linear regression models on out-of-sample transcripts for predicting downstream social and functional competency outcomes. Healthy controls were only evaluated on the SSPA task in part (a). The associated regression model performance statistics are available in Table 8.

4.3.3.2 Specific Level of Functioning Score Prediction

The SLOF assessment for all clinical participants (Sz/Sza and BD) was independently modeled for the three different subscales: (1) *interpersonal relations*, (2) *participation in home and community activities*, and (3) *work skills*. The three subscale scores are also summed into a composite functional competency score (*Fx composite*). Descriptive statistics for SLOF scores are shown in Table 7.

A summary of the cross-validation (for model development) and out-of-sample test performance for SLOF prediction models is shown in the bottom half of Table 8.

Table 8: This table shows the performance of the linear regression models developed to predict performance on *downstream* outcomes in social and functional competency, namely in the Social Skills Performance Assessment (SSPA) and Specific Level of Functioning (SLOF) tasks. All participants were evaluated on the SSPA task (from which the transcripts originated), but only clinical participants (those with Sz/Sza or BD) were evaluated for the SLOF tasks. The table shows the performance of the models in terms of coefficient of determination (R^2), the Pearson correlation coefficient (PCC), mean absolute error (MAE), and mean squared error (MSE) between the predicted and actual outcomes for each task for both the samples used for model development (cross-validation) and new unseen transcripts (out-of-sample).

		R^2	PCC	MAE	MSE
SSPA Avg.	Cross-validation	0.583	0.787	0.178	0.330
	Out-of-sample	0.611	0.785	0.171	0.330
SLOF:					
Interpersonal	Cross-validation	0.224	0.473	0.511	0.565
	Out-of-sample	0.298	0.569	0.493	0.545
Activites	Cross-validation	0.418	0.647	0.160	0.285
	Out-of-sample	0.298	0.569	0.493	0.545
Work	Cross-validation	0.286	0.535	0.734	0.728
	Out-of-sample	0.082	0.351	0.830	0.765
Fx Composite	Cross-validation	0.369	0.608	2.507	1.272
	Out-of-sample	0.358	0.616	2.422	1.238

4.3.4 Discussion

The primary goal of medical treatments, interventions, and therapeutics are to improve the quality of life for afflicted individuals. In order for treatments in schizophrenia, schizoaffective disorder, or bipolar disorder to be deemed successful, they must associate with measurable improvements in social and functional competency. For this reason, we believe it is critical to identify ways in which the computed language metrics described above can help clinicians understand downstream outcomes .

The downstream outcomes were largely split into two major categories: (1) social competency as measured by the SSPA task, and (2) functional competency as measured by the various subscales within the SLOF evaluation. We look at each one individually here.

4.3.4.1 SSPA

As the transcripts used to compute language features were derived directly from the SSPA task, it is unsurprising that the best predictions ascertained from the language samples are found in the regression analysis for predicting the average SSPA score from scenes 2 and 3. Using the same regression model for SSPA prediction, we see remarkably similar performance between the training samples in cross-validation and the out-of-sample test samples (to which the model designer did not have access during development).

The model used a combination of features to estimate the average SSPA score across the two scored tasks. The most important features domains in the model were *volition*, *lexical diversity*, *affect*, and *appropriateness of response*. These feature domains are known to be important for performance evaluation in the SSPA task, so we can be confident that the prediction model is working as intended and can be easily understood by those who may wish to use it in practice.

4.3.4.2 SLOF

The SLOF evaluation consists of three subscales for interpersonal relationships, participation in home and community activities, and work skills, with an additional

composite functional score. The results in Table 8 show that these variables were slightly harder to predict solely using computed features from the SSPA transcripts. However, the results between the cross-validation and out-of-sample participants were most consistent for *interpersonal relationships* and *activity participation*. *Work skills* were much harder to predict for out-of-sample participants using the constructed model; however, this is consistent with the fact that the SSPA task (with which the transcripts were collected) is most related to *interpersonal relationships* and *activities*. Hence, the data we collected lends itself best to predict performance on those subscales in real world functional competency tasks with the models we developed. Interestingly, the functional composite score prediction was very consistent for both the cross-validation and out-of-sample transcripts, showing that language samples collected on a limited task yield some predictive insight into functional competency as a whole. It is likely that a wider variety of language samples collected from tasks of a different nature could improve upon these predictions.

Chapter 5

REAL-WORLD IMPLEMENTATION CHALLENGES: ASSESSING THE IMPACT OF ERRORS IN AUTOMATIC SPEECH RECOGNITION

5.1 Investigating the Effect of ASR Errors on Sentence Embeddings (ICASSP 2019 Paper)

5.1.1 Introduction & Related Work

Many real-world applications motivate the need to accurately capture the semantic content of a sentence. Examples include sentiment analysis of product reviews, customer service chatbots, biomedical informatics, among several others. *Word embeddings* map words from a lexicon to a continuous vector space in which nearby vectors are also semantically related. Similarly, *sentence embeddings* map individual phrases or sentences to a continuous vector space that preserve the text semantics. The approaches to the word-embedding problem range from simple singular value decomposition of co-occurrence matrices [11] to neural network models trained on large corpora (*e.g. word2vec* [14], *GloVe* [15], and *FastText* [144]).

These approaches have revolutionized NLP research by showing impressive results on downstream NLP tasks; however, to the best of our knowledge, all of the previous work on sentence and word embeddings is built upon the assumption that the available text for training and testing each embedding model is perfectly transcribed. In most real-world applications, it is unlikely that textual language data will be free of error. In fact, an increasing number of applications rely on *automatic speech recognition* (ASR)

systems for transcriptions. The performance of an ASR system can be characterized by its *word-error rate* (WER), which defines the percentage of incorrect word errors given by the output of a particular system. Typical modern ASR systems have a WER ranging from ∼10% to ∼ 35% [145]. With a few exceptions, *i.e.* [146], [95], [94], [147], the effects of ASR errors have been largely ignored in many NLP applications. And, to the best of our knowledge, no previous work has been conducted to evaluate the effects of ASR errors on sentence embeddings and their performance in downstream NLP tasks.

In this work, we evaluate the robustness of several state-of-the-art sentence embeddings to word substitution errors typical of ASR systems[16]. To do this, we propose a new method for simulating realistic ASR transcription errors with a specified WER that is implemented with only publicly available tools for acoustic and semantic modeling. We evaluate the resultant embeddings on the semantic textual similarity (STS) task, a popular research topic in NLP within the area of statistical distributional semantics. In STS, the goal is to develop sentence embeddings that can successfully model the semantic similarity between two sentences (or another arbitrary collection of words). Several recently developed sentence embedding methods have shown very promising results on STS tasks [14], [15], [148], [74], [149], [76], [75], [150]; however, all have been evaluated using perfect transcripts. We attempt to re-evaluate the results on standard STS datasets after introducing the errors simulated using our approach. In short, the contributions of this work are: 1) a new simulator for introducing ASR-plausible word substitution errors that utilizes phonetic and semantic information to randomly replace words in a corpus with likely confusion words, 2)

[16]WER calculation includes unintended word *insertions*, *deletions*, and *substitutions*. We note that a limitation of our model is that it only considers potential substitution errors when simulating ASR error

Table 9: Example sentence pairs from STS-benchmark [152] and SICK corpora [151] after corrupting all sentences with WER of 30%. Substituted word errors are shown in italics. A high WER is used here to demonstrate the types of substitution errors simulated by our method, incorporating both semantic and phonemic distance measures.

Original Sentence	Corrupted Sentence
Obama holds out over Syria strike.	Obama *helps* out *every* *Sharia* strike.
Russia warns Ukraine against EU deal.	Russia warns *Euro* against EU deal.
Gov. Linda Lingle and members of her staff were at the Navy base and watched the launch.	Gov. *Cindy* Lingle *add* *mentors* of her *staffs* were at the *NASA* base and watched the *launcher*.
I have had the same problem.	*Eyes* have had the same *progress*.
A white cat looking out of a window.	A white cat *letting* out of a window.

an evaluation of five recent sentence embedding methods and their robustness to simulated ASR noise, and 3) an evaluation of the STS performance of these sentence embeddings with simulated ASR errors and a variable WER using the *SICK* [151] and *STS-benchmark* [152] datasets.

5.1.2 Word Substitution Error Simulation

In this section we propose a new word substitution error simulator intended to model plausible substitutions that an ASR algorithm might produce. Our approach is based on the observation that the nature of word substitution errors in ASR systems

depends on the phonemic distance between the true word and the substituted word (because of the underlying acoustic model) *and* on the semantic distance between the true word and the substituted word (because of the underlying language model). To that end, we define the probability of substituting word w_i with word w_j by

$$P_{\text{subs}}(w_j|w_i) = \alpha \cdot \exp(-\frac{d_{ij}}{\sigma^2}), \tag{5.1}$$

where d_{ij} is a notion of distance between w_i and w_j comprised of both the phonemic and semantic distance, σ is a user-defined parameter that controls the shape of the resulting probability mass function (PMF), and α is a normalization constant that makes the marginal PMF in Equation 5.1 sum to one for each given w_i.

5.1.2.1 Estimating the Substitution Probabilities:

Given a corpus for which we want to simulate word substitution errors, we first compute the set of all unique words. Next, we consider the pair-wise substitution error probabilities using Eqn. (5.1). Estimating the probability of a substitution requires that we estimate d_{ij}. Loosely speaking, we model the total distance as being comprised of a phonemic distance between the words (contribution of acoustic model in ASR) and a semantic distance between words (contribution of the language model in ASR).

To estimate the phonemic distance, we use a phonological edit distance between words w_i and w_j, d_{ij}^P [153], [154], [155], loosely based on the Levenshtein edit distance [156], which compares the number of single-character edits one string would need to be identical to another string. We consider ARPABET transcriptions based on the

CMU Pronouncing Dictionary [157] to similarly compute phonemic similarity. To encode each phoneme, we use the *articulation features* provided by Hayes in [158]. The result is a binary feature matrix for each English phoneme in ARPABET. The phonological edit distance between two words can be computed as the number of *single-feature* edits that are required to pronounce the first word like the second, as outlined by Sanders *et al.* in [153].

To estimate the semantic distance between the words, we use the *GloVe* embeddings [15] for every word in the corpus and estimate the pairwise *cosine distance* as

$$d_{ij}^{S} = 1 - \cos\theta_{ij} = 1 - \frac{\mathbf{w_1}^T \mathbf{w_2}}{\|\mathbf{w_1}\|_2 \|\mathbf{w_2}\|_2} \tag{5.2}$$

where $\mathbf{w_i}$ and $\mathbf{w_j}$ represent the vector representations of two distinct words w_i and w_j, and θ_{ij} represents the angle between the vectors.

5.1.2.2 Algorithm Implementation:

The total distance in Equation 5.1 can be modeled using some function of the two contributions discussed above, $d_{ij} = f(d_{ij}^{S}, d_{ij}^{P})$. However, this approach requires that we estimate the conditional probability in Equation 5.1 for every pair of words in a corpus; for large, realistic vocabulary sizes, this becomes prohibitively large.

To alleviate the need to estimate all pairwise probabilities, we only consider the $N = 1000$ semantically most similar words in the corpus using d_{ij}^{S} and estimate the marginal distribution for that subset of words, assuming that it is zero for all others. In addition, in Equation 5.1, we model d_{ij} using only the contribution from

the phonological edit distance. The parameter σ can be chosen and tuned based on empirical results. We found that setting σ equal to the average phonological edit distance between each cluster of potential replacement words and the target word provided reasonable results. The overall procedure is summarized in Algorithm 1.

Algorithm 1 Random replacement of words in a given a corpus with a specified WER to simulate realistic ASR errors.

1: **procedure** CORRUPT SENTENCES(corpus, WER)
2: Find all unique tokens, w_i, in the corpus that exist in the set of pre-trained *GloVe* embeddings
3: Filter all w_i to those in pronouncing dictionary
4: **for** each w_i **do**
5: Find w_j, $j = 1, \cdots, N$ most similar words by d_{ij}^S
6: ARPABET transcription for w_i, all w_j ▷ CMU Dict
7: **for** each w_j **do**
8: Compute d_{ij}^P from w_i to w_j, where $j = 1, \cdots, N$
9: Keep only M values of $d_{ij}^P \leq$ thresh, where $M \leq N$
10: **for** $j = 1, \cdots, M$ **do**
11: Compute $P_{\text{subs}}(w_j|w_i)$ ▷ Eq. 5.1
12: Randomly select words to replace given WER
13: Replace selected words with error words based on the probability distributions computed ▷ Line 11

In Table 9, we provide several examples of the substitution errors simulated at a given WER of 30%.

5.1.3 Sentence Embedding Methods

The sentence embedding methods described in this section have all been shown to perform well on STS tasks [159], [160] and serve as a representative set of models to evaluate robustness to ASR errors. A brief description of each method is provided below:

Simple Unweighted Average: A common sentence embedding implementation is a computation of the arithmetic mean for all word vectors that comprise a sentence. This serves as a simple but effective baseline with pre-trained *word2vec* embeddings [14]. Additionally, averages can be computed after removing stop words which contain little semantic content (*e.g.* "is", "the", *etc.*).

Smooth Inverse Frequency (SIF): Arora *et al.* propose SIF embeddings [74], which involve two major components. First, a weighted average of the form $\frac{a}{a+p(w)}$ is computed, in which a is a scalar value (a hyperparameter, tuned to 0.001) and $p(w)$ is the probability that a word appears in a given corpus. This weighting scheme de-emphasizes commonly used words (with high probability) and emphasizes low probability words that likely carry more semantic content. Additionally, SIF embeddings attempt to diminish the influence of semantically meaningless directions common to the whole corpus. To do so, all word vectors in a dataset are concatenated into a matrix from which the first principal component is removed from each weighted average.

Unsupervised Smooth Inverse Frequency (uSIF): Ethayarajh proposes a refinement to SIF known as uSIF, which claims improvements in many tasks (including STS) [150]. uSIF differs from SIF in that the hyperparameter a is directly computed (and not tuned), making it fully unsupervised. Additionally, the first m ($m = 5$) principal components, each weighted by the factor $\lambda_1, \cdots, \lambda_m$ are subtracted for the common component removal step. Here, $\lambda_i = \frac{\sigma_i^2}{\sum_{i=1}^{m} \sigma_i^2}$, where σ_i is the i-th singular value of the embedding matrix.

Low-Rank Subspace: Mu *et al.* propose a unique sentence embedding in which sentences are represented by an N-dimensional subspace rather than a single vector [149]. Given word vectors of dimension d and subspace rank of N, a sentence matrix is first constructed by concatenating word vectors and has dimension $d \times N$ (we use $d = 300$ and $N = 4$). Then, principal component analysis (PCA) is performed to identify the first N principal components whose span comprise a rank-N subspace in $\mathbb{R}^d$. We consider this method for our simulated ASR error analysis to test whether the subspace representation is more robust to ASR errors than a vector representation.

InferSent: Conneau *et al.* developed the *InferSent* encoder that utilizes a transfer learning approach [76]. The encoder is trained with a bidirectional LSTM neural network on the Stanford Natural Language Inference (SNLI) dataset, a labeled dataset that is designed for textual entailment tasks. The embeddings learned from the NLI task are then used to perform textual similarity tasks in STS.

Computing Similarities: Sentences represented by vectors (*i.e.* averages, SIF, uSIF, *InferSent*) can be compared with *cosine similarity*, closely related to d_{ij}^S in Equation 5.2. Cosine similarity is given as $\text{CosSim} = 1 - d_{ij}^S = \cos\theta_{ij} = \frac{\mathbf{w_1}^T \mathbf{w_2}}{\|\mathbf{w_1}\|_2 \|\mathbf{w_2}\|_2}$. For subspace similarity, the authors in [149] suggest the analogous concept of computing the *principal angle* between the rank-N subspaces for two sentences. This can be readily obtained from the singular value decomposition. If we let the matrices $U(s_1)$ and $U(s_2)$ have columns that each contain the first N principal components for sentences s_1 and s_2, the principal angle similarity given by:

$$\text{PrincAng}(s_1, s_2) = \sqrt{\sum_{t=1}^{N} \sigma_t^2} \tag{5.3}$$

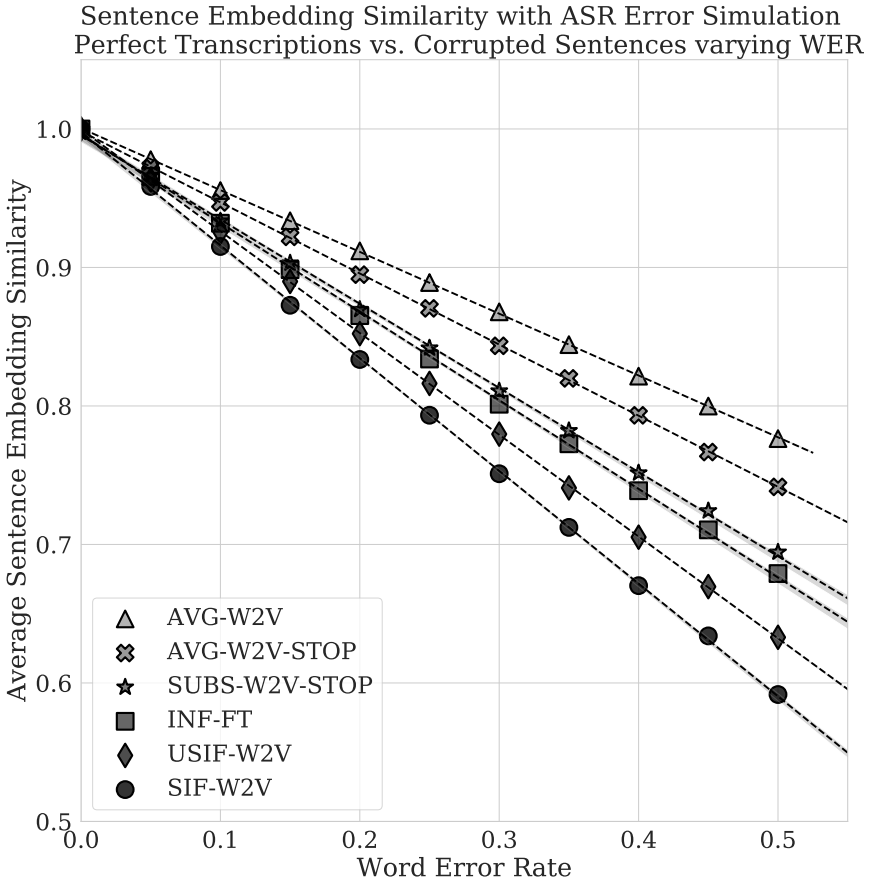


Figure 14: Regression plots for sentence embedding methods described in Section 5.1.3 as the WER is varied from 0% to 50%. We consider averaging *word2vec* vectors (△), averaging *word2vec* and removing stop words (𝕏), low-rank subspace representations with *word2vec* and stop-words removed (☆) [149], *InferSent* with *FastText* embeddings (□) [76], SIF with *word2vec* [74] (◯), and uSIF with *word2vec* (♢) [150].

In Equation 5.3, σ_t represents the t-th singular value of the product $U(s_1)^T U(s_2)$.

Sentence Embedding	**STS Corpus (dev & test set)**	**$PCC_{0\%}$ / $PCC_{10\%}$ / $PCC_{30\%}$ ($\times 100$)**	$\frac{PCC_{30\%}}{PCC_{0\%}}$
AVG-W2V:	SICK:	72.84 / 64.44 / 49.18	67.52%
	STS-benchmark:	67.40 / 59.23 / 45.64	67.72%
AVG-W2V-STOP:	SICK:	71.30 / 62.67 / 49.09	68.85%
	STS-benchmark:	68.61 / 62.15 / 49.99	72.85%
SIF-W2V:	SICK:	73.44 / 65.93 / 52.60	71.63%
	STS-benchmark:	70.39 / 63.51 / 52.06	73.96%
USIF-W2V:	SICK:	73.70 / 66.06 / 52.71	71.51%
	STS-benchmark:	69.95/ 62.85 / 51.11	73.07%
SUBS-W2V-STOP:	SICK:	66.10 / 59.28 / 46.94	71.02%
	STS-benchmark:	71.58 / 65.36 / 53.05	74.10%
INF-FT:	SICK:	75.94 / 68.95 / 56.56	74.48%
	STS-benchmark:	74.77 / 67.88 / 55.60	74.36%

Table 10: Pearson Correlation Coefficient (PCC) performance ($\times 100$) for SICK and STS-benchmark *dev* and *test* sets when WER is varied (0%, 10%, and 30%). The last column of each table shows the ratio (as a percentage) of the PCC at WER = 30% to the PCC at WER = 0% to demonstrate the robustness in STS performance of each sentence embedding to ASR errors at a high WER.

5.1.4 Results & Discussion

5.1.4.1 Robustness of Sentence Embeddings to Simulated ASR Errors

To study the effects of ASR errors on sentence embeddings, we first computed a sentence embedding for each sentence in SICK [151] and STS-benchmark [152] *dev* and *test* sets using each of the methods described in Section 5.1.3. Since *GloVe* embeddings were used to generate the simulated ASR substitution errors, we used *FastText* (for *InferSent*) and *word2vec* embeddings (all other methods) to generate sentence embeddings. For each method, we corrupted the sentences in the text with

a defined WER between 0% and 50% with the simulator described in Section 5.1.2. Then, each sentence in each set is compared with its corrupted counterpart using the relevant similarity metric (*i.e.* cosine or principal angle similarity).

The results are shown in Figure 14, in which all methods show a steady linear decline in average similarity between original and corrupted sentences as WER is increased. As expected, when WER is 0%, the sentence embedding similarity is equal to 1 for all methods. Simple averaging shows the least significant decline as WER is increased, *i.e.* at WER = 50% we see $\text{sim}_{\text{avg}} \approx 0.776$ for unweighted averaging and $\text{sim}_{\text{avg}} \approx 0.742$ for unweighted averaging and stop words removed. However, we see a significantly steeper decline for SIF and uSIF when WER = 50%, *i.e.* $\text{sim}_{\text{avg}} \approx 0.592$ for SIF and $\text{sim}_{\text{avg}} \approx 0.633$ for uSIF. The subspace representation and *InferSent* show a moderate decline in between these two extremes. These results are in line with our intuition, as we expect word substitution errors to have the smallest overall impact on unweighted average sentence embeddings. Also as expected, unweighted averages with stop words are more impacted by ASR errors, since stop words in the original corpus could be replaced by content words. This would lead to a greater difference between original and corrupted sentence similarity scores. SIF and uSIF are the most impacted by word substitution errors. We believe this is explained by the weighted average computation, *i.e.* if a frequent word is replaced by a less frequent word, it may have a greater impact on the overall sentence embedding. Additionally, it is likely the principal components of the embedding matrix are drastically altered by the introduced error and variance in the dataset, leading to larger differences in sentence embedding representations after corruption and common component removal. Since the common component removal is weighted by $\lambda_i \leq 1$ for each of the i principal

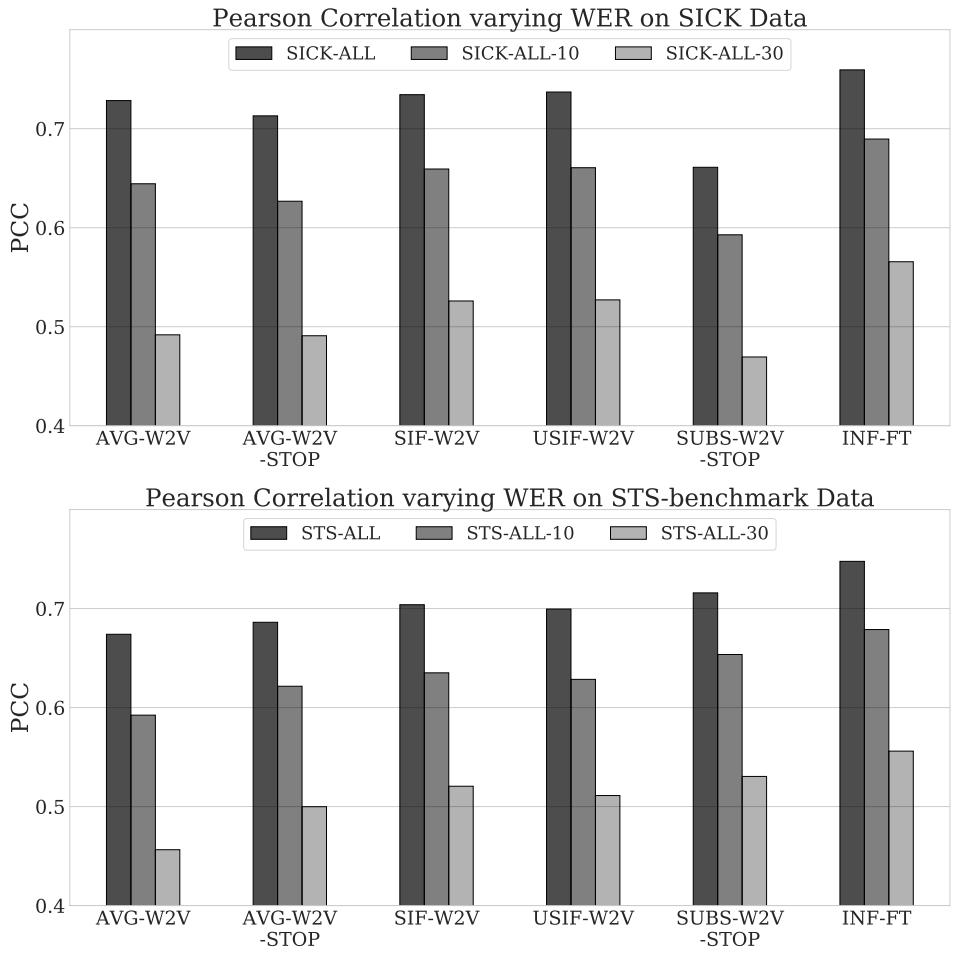


Figure 15: Graphical depiction of the STS performance of various sentence embeddings with simulated word substitution error, see Table 10

components in uSIF, the overall impact of the introduced variance due to ASR errors is diminished when compared to the single component removal step in SIF.

5.1.4.2 Evaluation of STS Results with Word Substitution Errors

We next compared the STS performance of the sentence embeddings on the original and corrupted corpora (with 10% and 30% WER) with the *dev* and *test* sets

of SICK [151] and STS-benchmark [152]. The *Pearson Correlation Coefficient* (PCC) between the computed similarities and the annotated similarity scores in the corpora is the standard metric by which we evaluate STS performance of a given method. The results are seen in Table 10 and Figure 15.

On the original sentences, simple unweighted averaging provides a strong benchmark for STS tasks on both corpora, with nearly equivalent results when stop words are removed. In most cases, the weighted average and de-noising provided by SIF and uSIF improve upon the results of unweighted averages, with both methods displaying near-identical performance. The subspace results are somewhat inconclusive, as they show a slight improvement over averages, SIF, and uSIF on STS-benchmark but a decrease in performance on SICK. The authors in [149] chose $N = 4$ empirically as the subspace rank, based on a variety of corpora which comprise the STS-benchmark set. It is possible that the absolute performance of the subspace sentence embedding can be improved by tuning the fixed subspace rank for SICK as well. Unsurprisingly, *InferSent* is consistently the strongest performer, likely due to its supervised training on the SNLI corpus.

When, ASR errors are introduced, the STS performance for each method changes significantly, as evidenced by the results in Table 10. Though the simple averages were least impacted with the introduction of ASR errors (Section 5.1.4.1), they perform worst among the methods tested on STS tasks with a high WER. On the other hand, SIF and uSIF embeddings were most impacted by ASR errors but perform among the best in STS when the WER is high. Again, we suspect this is due to the common component removal steps in SIF and uSIF, which effectively act as de-noising steps removing some of the additional variance in the embedding matrix due to substitution errors. Since SIF and uSIF display near-identical STS performance across both corpora,

we think uSIF may be a slightly better choice due to its increased robustness to ASR errors. Also, as suspected, we see that the subspace embeddings show increased STS performance robustness to word substitution errors when compared to averages if we consider the PCC ratio between high WER (30%) and original sentences. Subspace embeddings slightly outperform SIF and uSIF on STS-benchmark and slightly underperform SIF and uSIF on SICK by the same metric. Again, *InferSent* not only shows the best absolute performance on the original sentences, but shows the best performance with a high WER rate as well.

5.1.5 Conclusion

In this paper, we introduced a simulator that automates word substitution errors (given a WER) on perfectly transcribed corpora to simulate ASR-plausible errors, considering both phonemic and semantic similarities between words. We then used the simulator to intentionally corrupt standard corpora used for textual similarity tasks (SICK [151] and STS-benchmark [152]). From this, we were able to evaluate the impact that word substitution errors may have on some of the most recently developed techniques for sentence embeddings. We also evaluated the STS performance of each of these sentence embedding methods after introducing substitution errors with our simulator. We found several interesting results. For example, average sentence embeddings perform well for perfectly transcribed text, but show poorer STS performance when errors are introduced if compared to more advanced methods. On the other hand, pre-trained encoders, such as *InferSent* not only show state-of-the-art performance on STS tasks with perfectly transcribed text, but also seem to show increased robustness to error for STS performance. If it is not possible to use an

encoder like *InferSent*, the weighted average and smoothing provided by SIF/uSIF or the low-rank subspace representation by Mu *et al.* [149] seem to be reasonable improvements over simple averages when it comes to STS performance for high-WER transcriptions.

5.2 Investigating the Impact of Introduced ASR Errors for Clinical Predictions

5.2.1 Introduction

In the next set of experiments, we use Algorithm 1 defined in Section 5.1 to examine the impacts of plausible ASR errors on the experiments we discussed in Chapter 4 for our upstream and downstream predictions in clinical applications. For the experiments in Chapter 4, the transcripts used in our study were all manually transcribed from audio recordings by research specialists at Queen's University and provided to our group for computational analysis. However, we know that obtaining high-quality manual transcriptions of all clinical interviews can be burdensome in terms of manual labor and potentially very expensive, making this approach difficult to scale.

Therefore, it is safe to assume a certain amount of reliance on ASR systems to obtain transcripts for similar studies in the future; this leaves us with the important task of characterizing how ASR errors can negatively impact our modeling performance for important clinical predictions. In this section, we explore this concept more deeply with regard to our upstream classification experiments and downstream SSPA score regression model predictions.

5.2.2 Methods

In this set of experiments, we make use of the ASR error simulator introduced in Section 5.1 by applying Algorithm 1 to our set of 303 SSPA conversations from the participants whose demographics are summarized in Table 3. Each set of conversation transcripts from this data were then corrupted using the ASR error simulator with five specified word-error-rates (WERs); namely these are WERs of 5%, 10%, 25%, 40%, and 50%. While we know that typical WERs in ASR systems are 10-20%, we are interested to see the level at which clinical modeling performance degrades for several degrees of word substitution errors.

All raw features and principal components were re-computed for each corrupted version of the entire data set. Then, the principal components for each set of conversations were used to make the same upstream and downstream predictions we previously discussed in Chapter 4. For the sake of simplicity, we focus only on a subset of the upstream and downstream modeling cases previously discussed. Namely, these are the prediction of the average SSPA Score for the downstream application and the classification experiments for the upstream application. The models here are the independently developed ones with parameters as defined in Sections 4.2.1 and 4.3.2.

For each set of regression or classification experiments, we look at how the full set of evaluation metrics are impacted at each level of WER. In the SSPA prediction regression experiment, this includes the coefficient of determination (R^2), Pearson correlation coefficient (PCC), mean average error (MAE), and mean squared error (MSE) between the predicted and true SSPA scores. For the classification experiments, this includes the confusion matrices, area under curve (AUC) of the receiver operating

Table 11: Summary of results showing linear regression performance metrics for a wide range of word substitution error rates (WER) for prediction of average SSPA score using the model parameters defined in Section 4.3.2.

WER	R^2	PCC	MSE	MAE
0%	0.593	0.785	0.176	0.330
05%	0.547	0.757	0.196	0.346
10%	0.562	0.761	0.190	0.337
25%	0.544	0.740	0.198	0.343
40%	**0.409**	0.724	**0.256**	**0.384**
50%	**0.216**	0.689	**0.339**	**0.461**

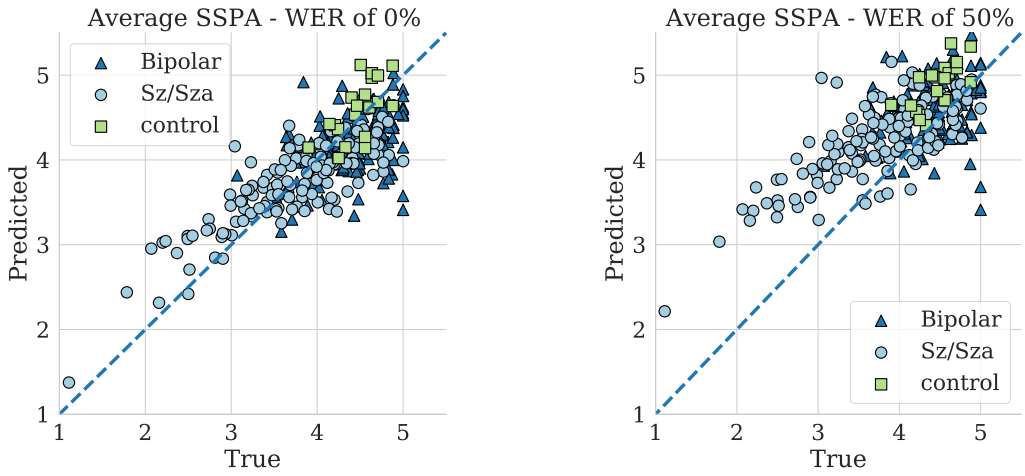
characteristic, and weighted average of the precision, recall, and $F1$ scores for each class.

5.2.3 Results

Here, we summarize how our previously developed models perform in the presence of noise with varying WERs for substitution errors.

5.2.3.1 Prediction of Average SSPA Score (Downstream)

Here we see the the performance metrics for predicting the average SSPA score for participants as the WER varies from 0% to 50%. The full set of performance metrics are provided in Table 11; visualizations of the best and worst case scenarios are then provided in Figure 16.



(a) 0% WER, $R^2 = 0.593$

(b) 50% WER, $R^2 = 0.216$

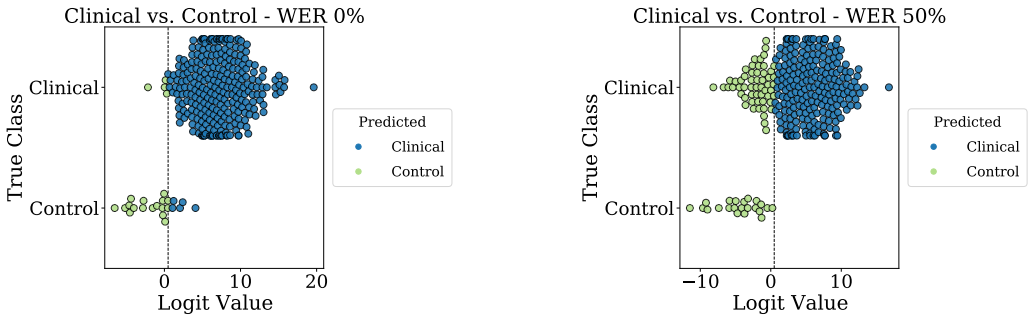
Figure 16: Visualization and comparison of best and worst cases of WER on SSPA score prediction. Table 11 shows the full results for all WERs.

Table 12: Summary of results showing logistic regression classifier performance metrics for a wide range of word substitution error rates (WERs) for clinical vs control classification model parameters defined in Section 4.2.1.

WER	Avg. Precision	Avg. Recall	Avg. $F1$
0%	0.896	0.879	0.887
5%	0.742	0.732	0.737
10%	0.904	0.746	0.803
25%	0.836	0.808	0.821
40%	0.718	0.910	0.773
50%	**0.643**	0.902	**0.668**

5.2.3.2 Diagnostic Classification Experiments (Upstream)

For the first (clinical vs. control) classification problem, The full set of performance metrics are provided in Table 12 as the WER varies; visualizations of the best and worst case scenarios are then provided in Figure 17.



(a) (b)

Figure 17: Visualization and comparison of best and worst cases of WER on clinical and control classification experiment. Table 12 shows the full results for all WERs.

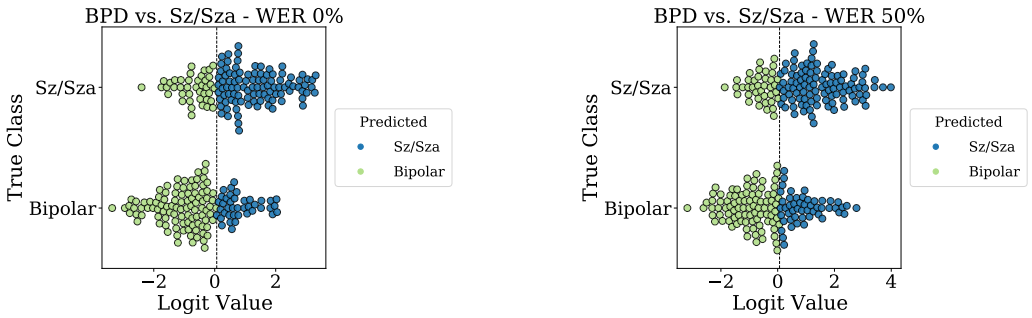
Table 13: Summary of results showing logistic regression classifier performance metrics for a wide range of word substitution error rates (WERs) for Sz/Sza vs BD classification model parameters defined in Section 4.2.1.

WER	Avg. Precision	Avg. Recall	Avg. $F1$
0%	0.712	0.712	0.712
5%	0.698	0.698	0.698
10%	0.726	0.723	0.721
25%	0.708	0.708	0.708
40%	0.698	0.698	0.698
50%	0.681	0.680	0.679

Similarly, the results for the BD vs Sz/Sza classification experiments with varying WER are provided in Table 13 and Figure 18.

5.2.4 Discussion

We can make several interesting observations about the impacts of speech recognition substitution errors on the modeling in our clinical SSPA experiments. Each case will be discussed individually.



(a) (b)

Figure 18: Visualization and comparison of best and worst cases of WER on Sz/Sza and BD classification experiment. Table 13 shows the full results for all WERs.

First, we observe the results from Table 11 and Figure 16 for the prediction of average SSPA score as WER increases. Unsurprisingly, the best performance in terms of the R^2 value for goodness of fit is for the uncorrupted version of all transcripts; however, interestingly, the performance hit does not seem to be particularly significant up until very high WERs of about 40% to 50%. This seems to indicate that our modeling approach is quite stable in the presence of noise until we reach an excessively high WER. A potential hypothesis for this may have to do with the fact that our measurement modeling approach that relies on PCA instead of low-level features for prediction may be more robust to the high variance associated with low-level features and lead to more consistently performing models; however, more thorough investigation comparing the two approaches would need to be conducted in order to confirm this theory.

Next, we can examine the same set of results for both of our classification experiments. The results for the clinical and control classification problem are outlined in Table 12 and Figure 17. Similar to the previous results we saw with the SSPA regression models, there is not a significant impact to model performance observed until the WER gets excessively high to around 40% to 50%; again this may be in-

dicative of the strength of the measurement model approach for clinical predictions. The corresponding results for the Sz/Sza and BD classification problem are found in Table 13, and Figure 18. Here, we notice that the impact of very high WER is still observable, though not nearly as pronounced. There are several potential reasons for this. First, as we discussed previously in Chapter 4, this is a significantly more complicated task to begin with, even in the absence of noise; therefore, the impact of introduced errors is more difficult to directly observe. Additionally, there are no healthy controls present in this experiment, meaning that the text that was being corrupted was likely to already be less coherent before being corrupted, suggesting that the impact of the introduced noise could be diminished. Again, more investigation here is needed.

5.2.4.1 Discussion of Positive Bias in Predictions

Another interesting observation in all of the above experiments has to do with how the corrupted predictions seem to be positively biased. In the regression experiments, the predictions at a 50% WER tend to be higher on average than those with a low WER; similarly, the corrupted transcripts are more likely to lead to a prediction of someone incorrectly being classified as a healthy control or an individual with BD rather than Sz/Sza. For the regression experiments, as WER increases, we are more likely to observe less correlation and increased variance in the predicted scores when compared to the true scores, but the R^2 value is most significantly impacted due to this bias. One reason for this may be the difference in raw word count for higher values of the WER parameter. We observed that as WER increases, we find that the raw word count of each transcript tends to increase as well. This may seem strange,

given that the error simulator only creates word substitutions and does not deal with insertions and deletions; however, the tokenizer used in our code may interpret some word substitutions (*e.g.* those with contractions) as multiple words, meaning that the word count could potentially increase for introduced errors. Since word count is such a strong factor in prediction of SSPA performance, the models may be more likely to assign higher scores or a greater likelihood of being a healthy control when it comes to transcripts with substitution errors using our method.

5.2.5 Conclusions and Next Steps

Clearly, these results show promise for the robustness of our measurement modeling approach for clinical prediction, but much more investigation is needed in order to quantify the degree to which this is true. These experiments would need to be repeated with several versions of corrupted transcripts to better characterize the impacts of error, and comparison studies would need to be conducted for predictions using low-level features compared with the principal components associated with the feature domains.

Additionally, it is clear that our simple ASR word substitution error simulator is limited in its capacity to realistically simulate ASR errors. Since, word count seems to play a big role, the impact of insertions and deletions will be critical to better characterize our understanding of the impacts of noise. One potential improvement here could also be to use a real ASR system in combination with a state-of-the-art text-to-speech (TTS) engine with our original conversation transcripts; this way, we may be able to see how well the simulator approximates ASR errors and conduct these experiments again to better understand the impact of noise.

Chapter 6

REAL-WORLD IMPLEMENTATION CHALLENGES: LACK OF AVAILABLE TRAINING DATA AND FULL AUTOMATION

6.1 Introduction, Background, and Motivation

As we have established, there are significant concerns regarding the repeatability and reliability of data driven methods for speech and natural language processing (NLP) for healthcare applications [122, 123, 121]. Due to these concerns, a myriad of challenges exist for developing machine learning models that can be widely adopted for use in clinical medicine; among these challenges are poorly labeled data, limited data, missing fields, privacy concerns, lack of standard formatting, and an under-representation of healthy individuals [161].

In our studies involving the SSPA transcripts [17, 22], we aimed to address some of the issues of interpretability and answering clinically meaningful questions using applications of machine learning. However, we also encountered the challenge of having an under-represented group of healthy controls to strengthen the findings of our studies. Several efforts have been made in simulating synthetic data for healthcare applications to ease concerns about patient privacy, but several adoption challenges still exist to verify that the synthetic data generation is representative of that of a true clinical sample for modeling purposes[162, 163].

To address these challenges, we explore the concepts of using *digital twin* (DT) technology [164, 165, 166, 167], generally defined as the virtual representation of a physical product or system that allows one to collect information or data that

is representative of the physical system it models. The DT pardigm has recently been explored in the context of improving healthcare technology, for example by using virtual representations of patients with historical data [168, 169]. Due to the imbalanced nature of our data (few healthy control samples), we aim to generate digital representations of patient interactions on the SSPA task that are reprsentative of the true sample of collected conversations in order to improve the robustness of our predictions models.

To do this, we take advantage of the state-of-the-art in langauge generation models, namely Generative Pretrained Transformer 3 (GPT-3), developed by OpenAI [170]. GPT-3 has shown immense promise on generating human-like text across many different scenarios using very few training examples as a prompt for the engine to generate additional text (few-shot learning). However, OpenAI also provides the opportunity to leverage GPT-3's largest models to fine-tuning towards a specific task given more data. We are therefore interested in addressing the following questions:

1. Can we effectively prompt GPT-3 to generate example conversations for our different groups of interest (Sz/Sza, BD, and healthy control) that are representative of the true samples?
2. Can GPT-3 provide us an interface in which it removes the human clinician from the loop in the data collection process?

Both of these questions provide us insight into the future opportunities of improving real-world language modeling in a clinical setting.

6.2 Methods

For the experiments here, we take advantage of OpenAI's GPT-3 application programming interface (API) [170]. GPT-3 is considered a few-shot learner in most cases, in which a user to provide a "prompt", with which it calls upon its engine to generate a response; the API provides several tunable parameters to set a maximum length of response, a stop sequence, randomness variables, and many others. However, for more specific use cases in which repeatable results are desirable, the API also provides a *fine-tuning* interface, in which users are able to create their own response generating engine by providing the system with several prompt and response pairs. Depending on the type of experiment we were doing, different fine-tuned GPT-3 engines were trained for each purpose, described below. From each of the generated conversation types, many of the same features that were previously computed (described in Chapter 3.2) are again computed in order to compare the quality of conversations provided by the AI engine to real human conversations.

6.2.1 Full Conversation Generation (Digital Twins)

The first set of experiments involved using the digital twin concept to use GPT-3 to generate full sets of conversations for all three of the SSPA scenes, first described in Chapter 4.2. Because healthcare data is generally very limited (and even more limited for healthy individuals), we are interested in augmenting our real-world data with virtual conversations; it is especially difficult to do collect this information at scale. In order to do this successfully, we must determine that the generated conversations match the parameters we compute, namely in terms of the feature distribution we see

from the real SSPA conversations for the groups of interest (Sz/Sza, BD, and healthy controls). The OpenAI API provides an interface for fine-tuning of the different provided GPT-3 engines, which vary in the number of parameters that make up the model. For all of these digital twin experiments, we use the *curie* engine for fine-tuning [170].

We used the following strategies for prompting the generation of full length conversations for each of the three SSPA scenes. Several approaches were examined for prompting GPT-3 in these fine-tuning experiments in order to examine whether the engine could serve as a digital twin to augment our data for the SSPA experiments. Each of these approaches involved using the full set of clinical and healthy control conversations used in our SSPA experiments described in the previous chapters. A brief overview of our approaches is described below.

6.2.1.1 Prompting by Participant Class

In the first explored approach, one model each was fine-tuned for each conversation scene of the SSPA experiments: (1) planning an activity with a friend, (2) introducing yourself to a new neighbor, and (3) negotiating with your landlord. For each model, we simply each of the participant types as a prompt to the GPT-3 fine-tuning API, *i.e.* schizophrenia/schizoaffective disorder (Sz/Sza), bipolar disorder (BD), and healthy controls.

Additionally, we also tried to see if GPT-3 was any better at capturing the intricacies of language if we separately fine-tuned a model to produce conversations for each participant type for a given contextual scene. This method would therefore

require a total of 9 fine-tuned GPT-3 engines, *i.e.* 3 fine-tuned models for each of the 3 scenes.

However, it was immediately clear that the prompts themselves were not very descriptive and the curie engine had difficulty understanding how to generate each of the different conversations. Therefore, we adopted several more strategies that involved using the actual SSPA scores as part of the prompt during the fine-tuning process.

6.2.1.2 Adding SSPA Scores to the Prompt

In the real SSPA transcripts, each individual was scored on a scale from 1 to 5 over many different dimensions of interest [21], from which an average score could be computed for each participant. Therefore, we believed it was possible to use the scores as part of the prompt with the associated real conversations as an expected response in the fine-tuning approach. Again, it became apparent there were several different methods with which this could be accomplished, with varying degrees of success.

The methods are summarized below:

1. We use "<*Participant group*>, <*Avg. SSPA score*>" as a prompt for fine-tuning with the real collected conversations, *i.e.* "*Sz/Sza*, 3.683"
    a) One GPT-3 model is fine-tuned for each of the three conversation scenes
    b) To generate the digital twin conversations, we prompt using a range of scores from 1.0 to 5.0, stepping by 0.5 for each of the three participant groups
2. Using the same fine-tuned models from (1), we now generate digital twin conversations by prompting the model for each scene with the same distribution

of SSPA scores that were found in each of the three participant groups for our real conversations.

3. Lastly, we attempt to give a descriptive prompt based on the SSPA performance for each individual as the training prompt for fine-tuning with real conversations. For example, for Scene 3 (landlord negotiation), we would say the following:

   > *This is a conversation between a tenant and a landlord. The tenant is upset about a leak in their apartment. The landlord is hesitant and unwilling to take care of the problem. The tenant is an individual with* <schizophrenia, bipolar disorder> *and has* <little, some, a lot> *of willingness to discuss the problem. This results in a* <short, medium length, long> *conversation. The tenant received a score of* <score> *out of 5.0 for this interaction.*

   a) For each scene, one model is fine-tuned for each participant type.
   b) The number of conversations generated match the scores and distributions for the the real conversations.

We then evaluate each generated conversation using the same analytical tools we described previously in Chapter 3.2.

6.2.2 Conversational Chatbot Experiments

In the next set of experiments, we focus on leveraging the power of generative language modeling with GPT-3 to aid us in data collection. As stated previously, one big challenge we face in our work is the relative scarcity of clinical assessments of healthy individuals as controls for these studies.

One potential bottleneck that makes data collection difficult is the need to have a trained clinician to administer the test to an individual who must be physically present. However, GPT-3's contextual language capability provides us an opportunity to collect potentially viable assessments of an individual's social skills by removing the human assessor from the process and instead having individuals speak directly to a fine-tuned GPT-3 chatbot engine.

For this set of experiments, we fine-tuned two sets of GPT-3 models for Scenes 2 (new neighbor) and 3 (landlord) of the SSPA task using the real human conversations as prompt-response pairs; one set of models for each scene was based on the *curie* engine while the other was based on the larger *davinci* engine.

Table 14 shows the strategy for fine-tuning a GPT-3 engine to create a chatbot model for a sample conversation transcript. In Table 14a, we see a partial transcript of a Scene 3 landlord conversation with "S" representing the participant and "A" representing the clinical assessor, whom we would like to replace with the chatbot in our experiments. Then in Table 14b, we see how the conversation is then formatted to be interpreted by the GPT-3 fine-tuning API. This strategy is repeated for all *healthy control* conversations we had in our sample (a total of $n = 22$).

Several individuals were then recruited to interact with all trained versions of the chatbot. Each had multiple conversations, and the ones that were deemed to be successful interactions (*i.e.* comparable to speaking to a human in the same context) were saved for further analysis.

Table 14: Fine-tuning strategy for chatbot training

(a) Sample conversation on which the GPT-3 model is fine-tuned.

Partial Transcript	
A:	Hello, this is Mr. Jones the Landlord.
S:	Hi, Mr. Jones! I am okay, but I have a problem with the leak that I called you about in my apartment last week.
A:	I'm sorry I haven't been over to fix it, I have been very busy.
S:	I understand, but the problem is getting pretty bad ...
A:	It didn't sound that bad the last time you called.
S:	···
	⋮

(b) Each response from the clinical assessor (A) is given to GPT-3 as a desired response with the entirety of the conversation preceding that response as the prompt for fine-tuning the chatbot responses.

	Prompt		Response	
1:	-		A:	Hello, this is Mr. Jones the Landlord.
2:	A:	Hello, this is Mr. Jones the Landlord.	A:	I'm sorry I haven't been over to ...
	S:	Hi, Mr. Jones! I am okay, but ...		
3:	A:	Hello, this is Mr. Jones the Landlord.	A:	It didn't sound that bad last time ...
	S:	Hi, Mr. Jones! I am okay, but ...		
	A:	I'm sorry I haven't been over to ...		
	S:	I understand, but the problem is getting pretty bad ...		
4:	A:	Hello ...	A:	···
	S:	···		
	⋮		⋮	

6.2.2.1 Performance Evaluation

To evaluate performance, a full set of features across seven feature domains (volition, affect, sentiment, syntactic complexity, semantic coherence, and appropriateness of response) were computed just as they were described in Section 3.2. These features are then compared with the similarly computed features on the real conversations (with healthy controls) to see if similar upstream and downstream predictions can be made using our previously developed models.

We also trained new linear regression models for the prediction of SSPA scores in Scenes 2 (new neighbor) and 3 (landlord) using the full set of previously collected

conversations with real individuals. While we are only interested in the healthy control conversations when comparing with our set of healthy individuals conversing with the chatbot, it was helpful to use all types of conversations when training our regression models to have an abundance of data and a wide range of computed metrics and scores to predict. In total we have $n = 141$ sets of conversations for those with Sz/Sza, $n = 140$ for those with BPD, and $n = 22$ healthy controls. A stratified split of the full data set was conducted such that 20% of the individuals in each group were held out and not used during the training process. Then, the remaining transcripts were used individually for Scenes 2 and 3 to train linear regression models using the average SSPA score for each scene as the target prediction value.

The model was fit with elastic net regularization, a regularization method to avoid overfitting that combines both LASSO (ℓ_1-norm) and ridge (ℓ_2-norm) regularization methods. Equation (6.1) gives the definition of the loss function which is minimized for this form of linear regression with elastic net regularization, implemented with the *scikit-learn* toolbox in Python [171].

$$\mathcal{L} = \frac{1}{2 * n_{\text{samples}}} * ||y - Xw||_2^2 + \alpha * \ell_1\text{-ratio} * ||w||_1 + \alpha * \frac{(1 - \ell_1\text{-ratio})}{2} * ||w||_2^2 \quad (6.1)$$

In Equation (6.1), the goal is to learn the values of w that minimize the loss $\mathcal{L}$ that allow us to best predict the target values y from X. The regularization hyperparameters in this case are α and the ℓ_1-ratio. To optimize their choice, we performed leave-one-out cross validation (LOOCV) with the training transcript set over a wide range of choices for α and ℓ_1-ratio, and determined the optimal values that led to the best fit according to the LOOCV.

Once the prediction models had been trained for both the Scene 2 and Scene 3 SSPA scores, they are validated by observing their performance on the held out test

data set. Lastly, we use the trained model parameters to then predict the SSPA scores for the collected chatbot conversations. Since the chatbot users did not have a real clinical SSPA evaluation, these were validated by comparing the distribution of predicted scores from the healthy control individuals who spoke with real humans.

6.3 Results

6.3.1 Full Conversation Generation - Digital Twins

Because so many prompting strategies were used to generate digital twin conversations with GPT-3 (described in Section 6.2.1), we had a large number of conversations that were available for determining if the GPT-3 *curie* engine was capable of capturing the linguistic features of interest when generating these conversations using varying prompts that matched the real ones.

We found that the last strategy mentioned in Item 2 in Section 6.2.1.2, *i.e.* using a prompt consisting of the participant class and SSPA score to generate each conversation which matched the distribution of real scores seen in our true sample. Here, we focus on this set of generated conversations for SSPA Scenes 2 and 3, the new neighbor and landlord negotiation tasks.

As previously mentioned, we fine-tuned a GPT-3 model for each type of participant individually with this strategy, which yielded the best results we found from all of our strategies. However, since healthy control data was limited even in the real conversations, we did not have a sufficient sample with which to fine-tune a healthy control GPT-3 conversation-generating model. Therefore, for these experiments, we

Table 15: Summary of the number of conversations that were deemed to be of good quality (GREEN), average quality (YELLOW), or poor quality/unusable (RED) based on manual inspection for digital twin experiments.

Participant Class	Scene	Good	Average	Poor	**Total**
Bipolar	*Sc.* 2	113	15	9	**137**
	Sc. 3	114	12	10	**136**
Sz/Sza	*Sc.* 2	109	17	11	**137**
	Sc. 3	112	16	10	**138**

report our findings for conversations that were generated matching the real distribution of Sz/Sza and BD conversations for Scenes 2 and 3 of the SSPA.

6.3.1.1 Conversation Quality

While GPT-3 is very powerful, results with the same prompt are often not repeatable. Therefore, for this step, each generated conversation had to be manually evaluated for quality, as some generated results could get stuck in a repetitive loop, output nonsensical text, or be completely blank. The summary of the quality of the generated conversations is found in Table 15. Conversations categorized as good quality (GREEN) were mostly or entirely successfully generated, with both individuals interacting in a similar manner to the true human sample. Average quality (YELLOW) conversations were identified as those in which some problems existed, *i.e.* lines that did not make sense or repetitive, but still contained much usable material. Poor quality (RED) conversations were those that were completely unusable due to being blank or completely nonsensical or repetitive.

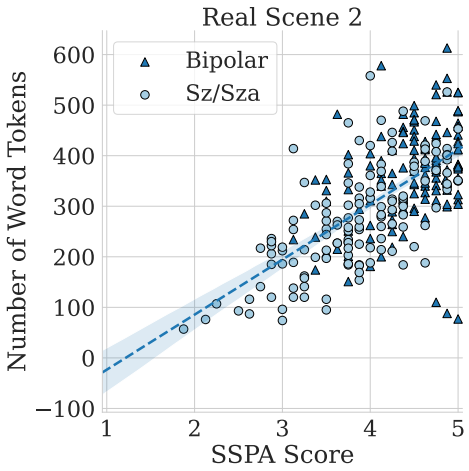
From the results in Table 15, we see that approximately 81% of the generated conversations in each case produce readable results that look similar to the real

conversations, approximately 11% are mostly usable with some issues, and about 8% are completely unusable. From these, we can compute the features on the conversations marked GREEN and YELLOW and compare them with the features computed on the original human conversations to see if the fine-tuned GPT-3 *curie* engine is capturing the linguistic dimensions of interest for these clinical groups.

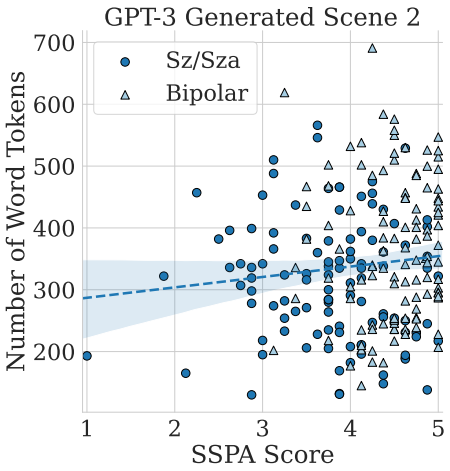
6.3.1.2 Comparing Generated and Real Conversations

In order to maximize the probability that our generated conversations remember the real ones, we only consider the ones that were deemed to be of good quality (GREEN) in Table 15.

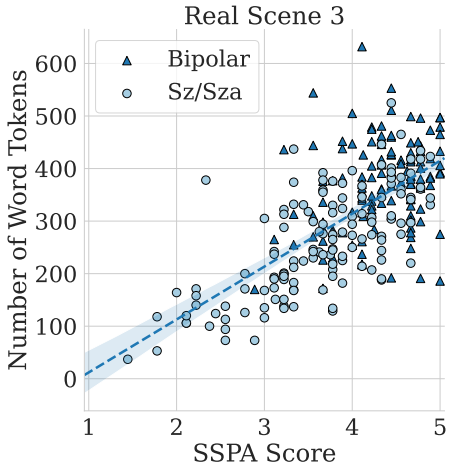
From our previous work, we know that the verbal output, *i.e.* the number of word tokens in each conversation, is highly correlated with performance on the SSPA task. Therefore, we are interested to see if the same is true for the SSPA conversations for Scenes 2 and 3 with the digital twin conversations generated by GPT-3. These results are all summarized in Figure 19. As we can see, there is far less correlation (if any) in both scenes between verbal output and the the associated score on the SSPA assessment when conversations are generated by our digital twin model. Since all of the predictive modeling tasks described in Chapter 4 rely on the importance of word count for their predictive ability, we can say with some certainty that the digital twin model here is not truly representative of the real test sample.



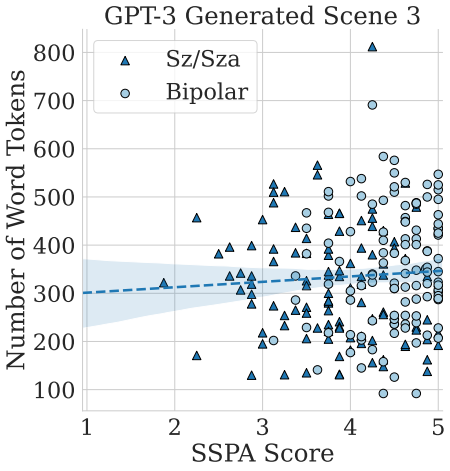
(a) Pearson $R = 0.649$



(b) Pearson $R = 0.106$



(c) Pearson $R = 0.670$



(d) Pearson $R = 0.062$

Figure 19: Comparison of the verbal output (in number of word tokens spoken) between the real human conversations and digital twin conversations using GPT-3 generation for SSPA Scenes 2 and 3.

Table 16: Number of conversations and unique participants for each SSPA interaction.

	GPT-3 Engine	Unique Participants	Total Conversations
New Neighbor (Sc. 2)	*curie*	6	15
Landlord (Sc. 3)	*davinci*	6	25

6.3.2 Conversational Chatbot Experiments

Next, we look at the results from the conversational chatbot experiments, for which we collected conversations from multiple individuals interacting with GPT-3 playing the role of the clinician in both SSPA Scenes 2 (new neighbor) and 3 (landlord). Here, our goal is to see how representative these conversations are when compared with the healthy control conversations from the real SSPA tasks.

6.3.2.1 Summary of Collected Conversations

One GPT-3 model was fine-tuned for each of the SSPA scenes. Qualitatively, we found that the *curie* engine performed best for the landlord conversations, whereas the *davinci* engine provided the best results for the new neighbor conversations. A summary of successful interactions (determined by the user in each case) can be found in Table 16.

6.3.2.2 Comparison of Feature Distributions

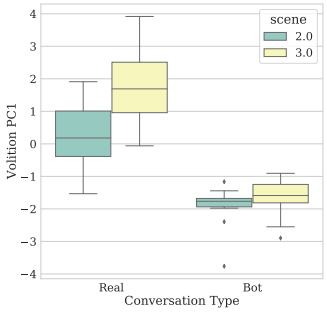
All of the features we describe in Chapter 3.2 are computed on the transcripts of the conversations each user had with the chatbot for both the new neighbor and

landlord interaction scenes. Here, we compare the distribution of features when a human clinician is used to collect the samples against those that are collected by GPT-3 playing the role of the clinician.

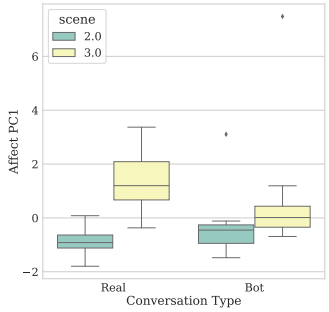
Prinicpal components for the seven feature domains described in 3.2 were computed for the real human conversations, preserving 85% of the variance in the data for each domain. Then, in order to compute the same PCs for the chatbot conversations, we used the derived projection coefficients from the real conversations. In order to compare both, statistical testing was performed to identify the probability that feature distributions for conversations collected with the chatbot followed a similar distribution to those collected by humans. Here, we use a two-sample Kolmogorov-Smirnov (KS) test [172]; in this case, the null hypothesis is that the features for the chatbot and the human conversations both arise from the same distribution.

In many cases, we end up rejecting this null hypothesis with a very high probability. In fact, the feature distributions for the chatbot and human conversations are drastically different even upon visualization in most cases. The results of the KS-test for the distribution of principal components across all feature domains is found in Table 17. However, a few feature domains do have similar looking distributions; in particular, we see that for lexical density, semantic coherence, and appropriateness of response, many of the distributions look similar between the chatbot and real conversations.

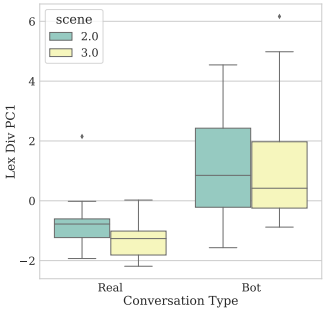
A few example box plots of the distributions of features for the Scene 2 and Scene 3 conversations are seen in Figure 20.



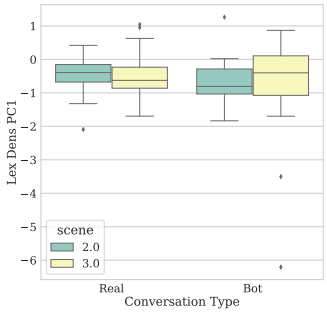
(a) Volition PC1



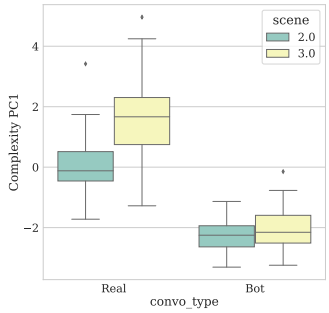
(b) Affect PC1



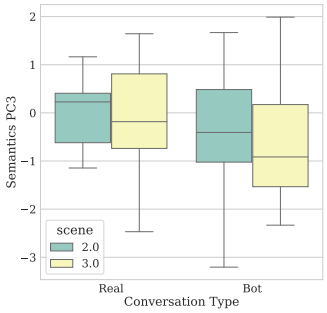
(c) Lexical Diversity PC1



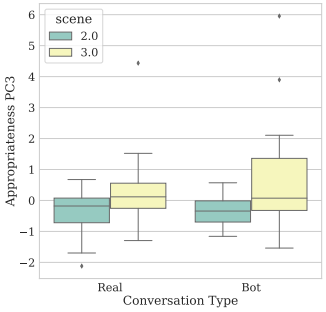
(d) Lexical Density PC1



(e) Syntactic Complexity PC1



(f) Semantic Coherence PC3



(g) Appropriateness of Response PC3

Figure 20: Box plots showing selected feature PC distributions for the seven feature domains. Results are shown individually for features computed on Scene 2 and Scene 3 conversations for both the real conversations and those that were conducted using the GPT-3 chatbot.

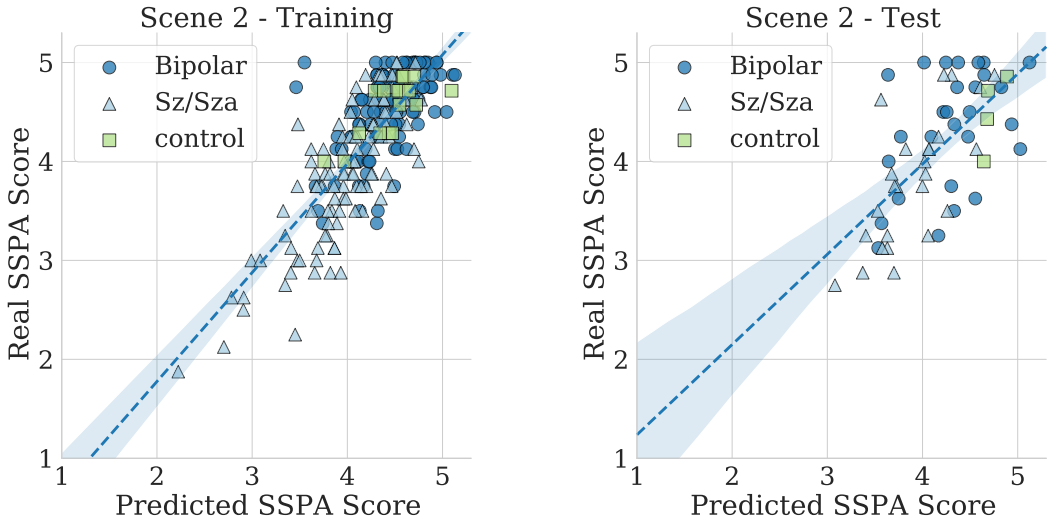
Table 17: Results of the Kolmogorov-Smirnov test for the principal components for each of the feature domains when comparing the real human conversations to the conversations collected by the GPT-3 chatbot.

	Scene 2: KS-Statistic	p-value	Scene 3: KS-Statistic	p-value
Volition PC1	0.95455	0.00000	1.000000	0.00000
Volition PC2	0.45455	0.02894	0.45455	0.02005
Affect PC1	0.44318	0.03633	0.68182	0.00004
Affect PC2	0.57955	0.00195	0.68182	0.00004
Lex Div PC1	0.67614	0.00015	0.77273	0.00000
Lex Div PC2	0.89205	0.00000	0.90909	0.00000
Lex Dens PC1	**0.38068**	**0.10132**	**0.18182**	**0.87168**
Lex Dens PC2	**0.19886**	**0.77299**	**0.27273**	**0.39374**
Complexity PC1	0.89205	0.00000	0.90909	0.00000
Semantics PC1	0.67614	0.00015	1.00000	0.00000
Semantics PC2	**0.34659**	**0.16858**	0.68182	0.00004
Semantics PC3	**0.27841**	**0.38382**	**0.36364**	**0.10926**
Semantics PC4	**0.33523**	**0.19261**	**0.59091**	**0.00067**
Appropriateness PC1	**0.41477**	**0.05892**	0.77273	0.00000
Appropriateness PC2	**0.35795**	**0.14192**	0.50000	0.00729
Appropriateness PC3	**0.23295**	**0.59851**	**0.27273**	**0.39374**
Appropriateness PC4	**0.40341**	**0.07109**	0.54545	0.00236

6.3.2.3 Comparison of SSPA Model Predictions

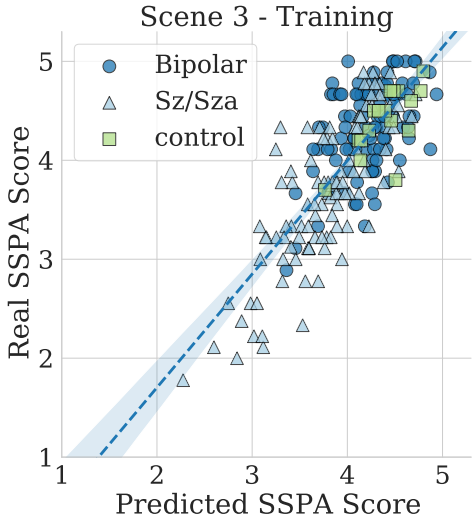
In the next set of experiments, we developed prediction models to predict the average SSPA score individually for Scenes 2 and 3, in a similar to the method followed in for the models previously shown in Section 4.3. Here, we again used the real human conversations to determine the PCA projections that we used to compute the feature components for the chatbot conversations.

Then, using all the PCA and raw features, prediction models were trained to predict the individual SSPA Scene 2 and Scene 3 scores with elastic net regularization. The results from the final models are shown in Figure 21. As opposed to what we did

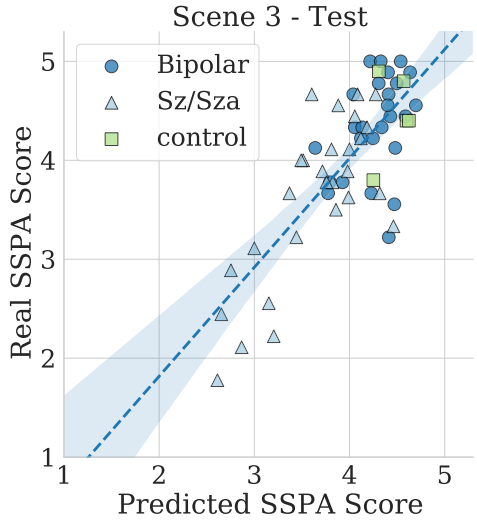


(a) $R^2 = 0.630$, MSE $= 0.148$, PCC $= 0.797$

(b) $R^2 = 0.422$, MSE $= 0.248$, PCC $= 0.656$



(c) $R^2 = 0.606$, MSE $= 0.181$, PCC $= 0.785$



(d) $R^2 = 0.582$, MSE $= 0.235$, PCC $= 0.766$

Figure 21: Linear regression fit with elastic net regularization using a random stratified split of the real SSPA conversation transcripts collected from 203 individuals. Models were trained using ALL raw features and principal components for each feature domain. Results for goodness of fit are shown with coefficient of determination R^2, mean-squared error (MSE), and Pearson correlation coefficient (PCC).

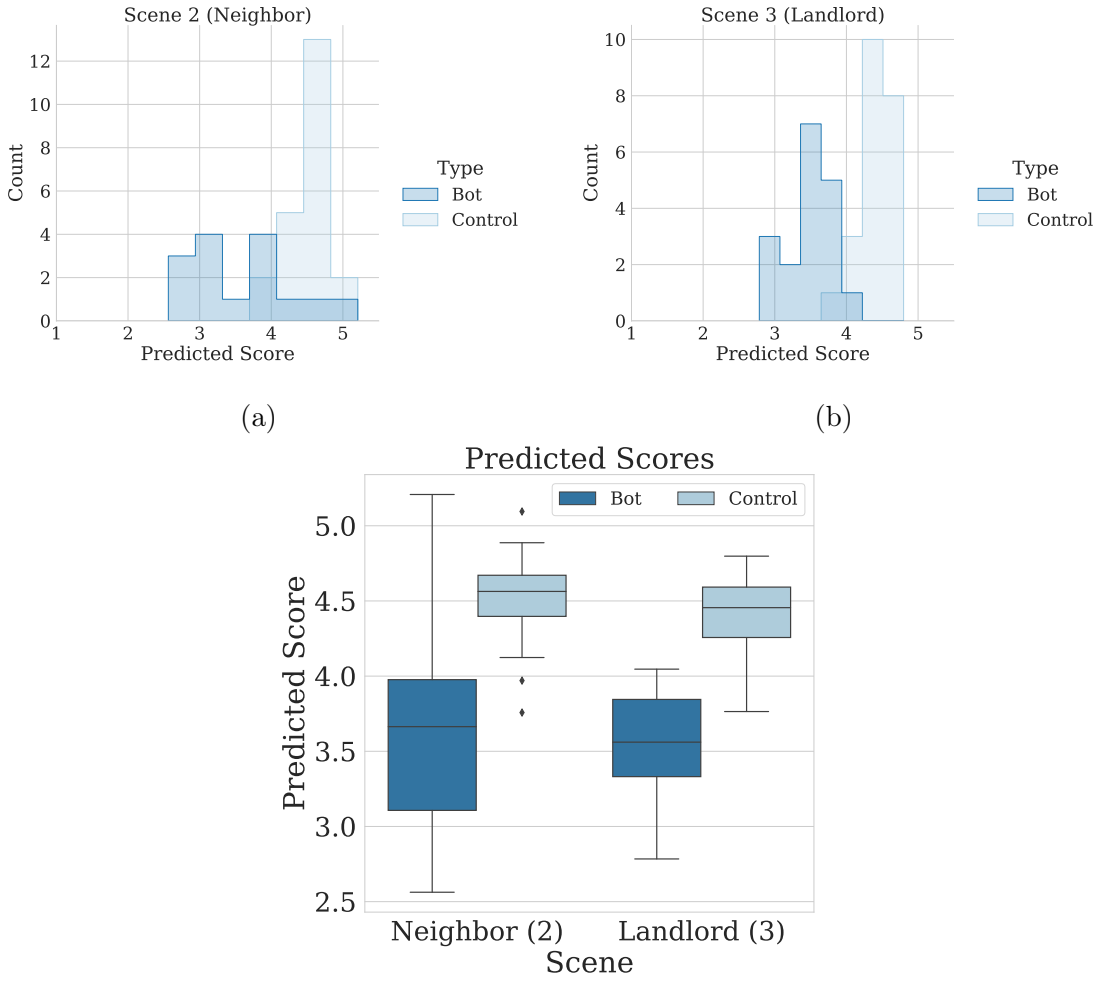
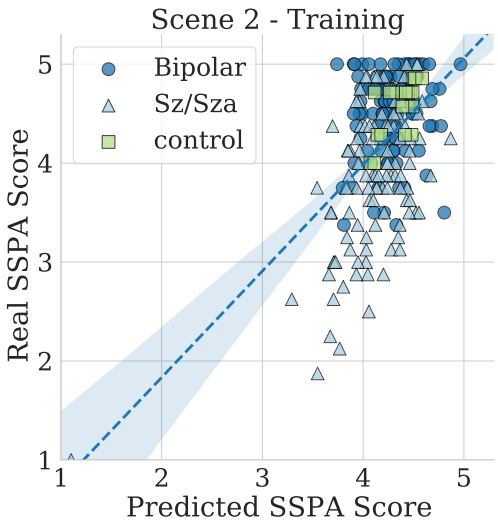
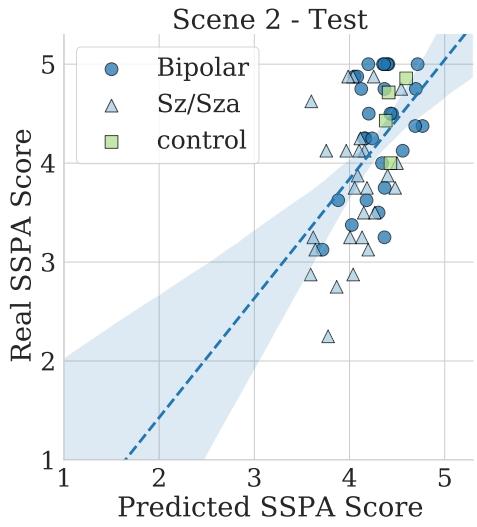


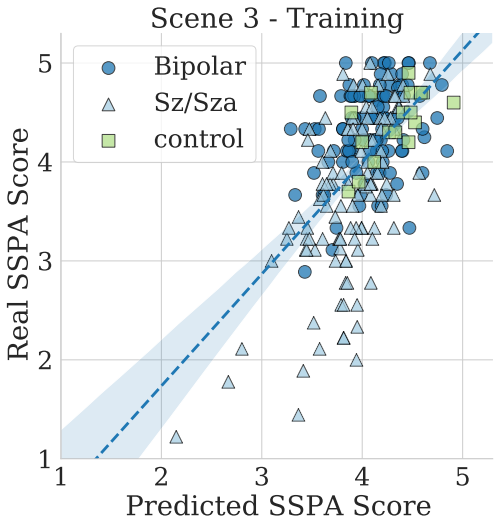
Figure 22: Trained models using ALL raw features and principal components. We see histogram and box plot representations of the difference in distributions for model predictions of SSPA score for the real control conversations and chatbot conversations, separately for Scenes 2 and 3 of the SSPA.



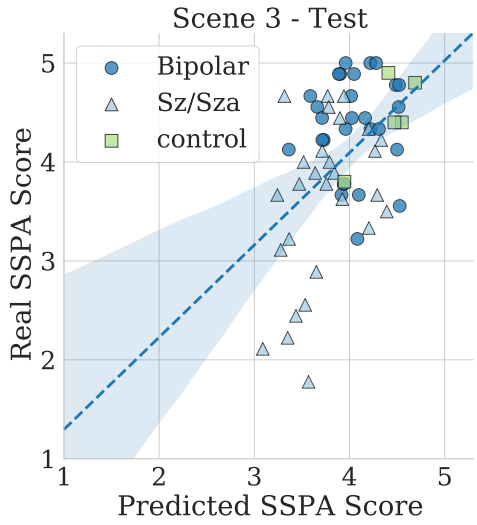
(a) $R^2 = 0.328$, MSE $= 0.352$, PCC $= 0.536$



(b) $R^2 = 0.206$, MSE $= 0.380$, PCC $= 0.493$



(c) $R^2 = 0.328$, MSE $= 0.352$, PCC $= 0.576$



(d) $R^2 = 0.212$, MSE $= 0.440$, PCC $= 0.480$

Figure 23: Linear regression fit with elastic net regularization using a random stratified split of the real SSPA conversation transcripts collected from 203 individuals. Models were trained using only the principal components which had similar distributions for both the real healthy control conversations and the chatbot conversations (see bold items in Table 17). Results for goodness of fit are shown with coefficient of determination R^2, mean-squared error (MSE), and Pearson correlation coefficient (PCC).

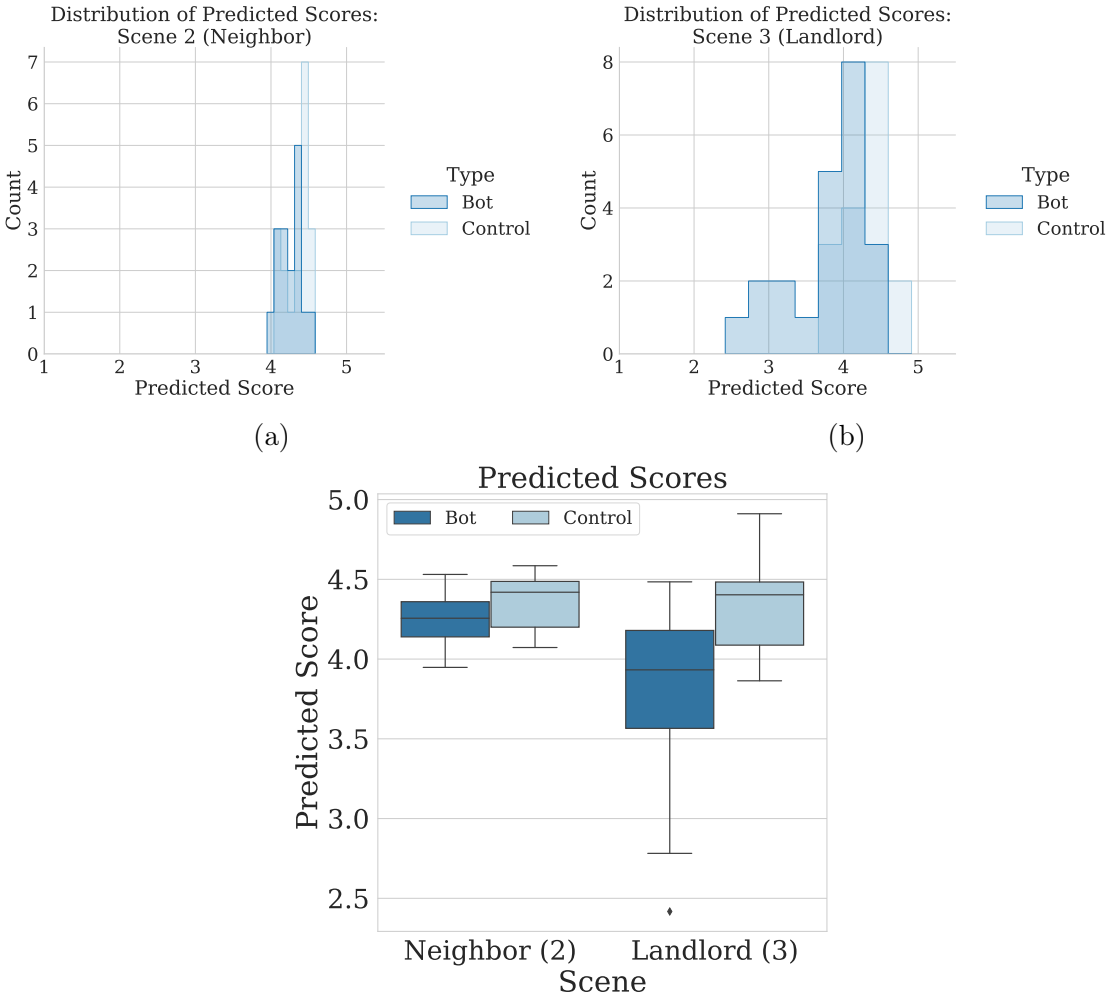


Figure 24: Trained models using only the principal components which had similar distributions for both the real healthy control conversations and the chatbot conversations (see bold items in Table 17). We see histogram and box plot representations of the difference in distributions for model predictions of SSPA score for the real control conversations and chatbot conversations, separately for Scenes 2 and 3 of the SSPA.

in 4.3, we now need individual models for Scene 2 and Scene 3; averages do not suffice since the same individuals did not converse with the landlord and neighbor chatbot models.

Next, we used these trained models to predict SSPA scores based on the same features computed from the chatbot conversations. Since these individual conversations were not scored in a clinical setting, we do not have true expected values for these predictions. Since all of our test subjects were healthy individuals, we determined that we could compare the values of these predictions with those for the healthy control individuals in the original study, as a measure of how similar the chatbot conversations are to the in-person conversations. In Figure 22, we see some comparisons of the distributions of these predictions in histogram and box-plot form.

6.4 Discussion

6.4.1 Full Conversation Generation - Digital Twins

The digital twin experiments for generating full conversations provided us with significant insight to both the capabilities and limitations for using GPT-3 to augment our training data for the development of robust language models to be used in real-world clinical settings.

Largely, we found that it was difficult for the GPT-3 *curie* engine to generate conversations which contained the important subtle differences in language among our groups of interest in a way that mirrored the true conversations. While some similarities in the distribution of verbal output were observed, we typically saw a much wider range of values for all features in the GPT-3 conversations across all

groups; this was true even after manually inspecting all the generated conversations and removing ones that were blank, nonsensical, stuck, or had other issues.

Hence, we can conclude that GPT-3 is not quite ready to to understand and pick up on all of the subtle linguistic differences that are present in the real population based on these limited prompting scenarios. This is likely due to the fact that GPT-3 is not trained on clinical speech samples but instead on publicly available online text.

For this reason, we are especially reliant on at least having one human participant in the loop for the collection of data with which we can train more effective and robust models for clinical applications.

6.4.2 Conversational Chatbot Experiments

Given that the digital twin generated conversations were not very representative of our true sample, we moved on to the conversational chatbot experiments with human participants interacting with the trained chatbot models.

From the results, we see that certain feature distributions (*i.e.* the PCs associated with appropriateness of response) actually do somewhat resemble the features computed from real human conversations; in particular, lexical density, semantic coherence, and appropriateness of response all have similar distributions when comparing the chatbot and real conversations. However, this is still largely not true for many of the features, especially those regarding volition and syntactic complexity.

There are several potential reasons for these discrepancies, largely due to the difference in the way individuals interact with chatbots as opposed to with other humans. When looking at the distribution of the feature domains (Figure 20 and Table 17, we immediately observe a large discrepancy in the Scene 2 and Scene 3

feature distributions for the volition components (Figure 20a). The verbal output of users tends to be much more concise when conversing with a chatbot by typing through a computer screen. Additionally, users tend to make several natural pauses and ramble through sentences when speaking spontaneously in person, possibly even more so in a role-playing exercise where they assume an unfamiliar persona. Instead, while typing to the chatbot, users can take their time to craft a concise response with a completely formed thought. For these reasons, we see that the user's sentence structures between the real-life and chatbot conversations are drastically different; this is also apparent in the distributions for Complexity PC1 (Figure 20e), since the sentence structures of chatbot conversations tend to be more concise with fewer embedded clauses.

Next, we trained new prediction models to individually predict the Scene 2 and Scene 3 SSPA scores using all 203 original transcripts from the three groups (Sz/Sza, BPD, and healthy controls). The first approach involved using all of the raw features and principal components for the seven feature domains as inputs to develop the prediction. Looking at the training and test set fits with elastic net regularization (Figure 21), we see that the test set performance is slightly worse but comparable to the training set prediction accuracy. We can conclude, as we did in our original study, that the SSPA scores can be reasonably well predicted with new real conversation transcripts.

In order to evaluate the chatbot approach, we looked at the distribution of predicted SSPA scores using the same models with the chatbot conversations and compared them to the full set of 22 healthy control conversation prediction scores. The results are shown in Figure 22. Upon inspection of the histograms and box plot for the predicted score distributions, it is clear that the real healthy control conversations

and chatbot conversations have drastically different score distributions; we see that the real control conversations are tightly distributed near the high end of the SSPA score range, while the chatbot conversations have a much wider distribution. In our original models, we found the raw word count (W) and other measures of volition, measures of lexical diversity, and many others to be significant predictors for SSPA performance. However, we know that many of these features are distributed quite differently when comparing the chatbot and real conversations, and we therefore expect the model performance to be significantly worse at assessing the social skills of healthy individuals speaking to the chatbot.

Therefore, one final step was taken to re-train the models using only the PCA domains that we found to be similarly distributed between the human and chatbot conversations for SSPA Scenes 2 and 3; these are seen as the bold items in Table 17. Each prediction model was re-trained using only these features as inputs, and the results are shown in Figure 23. Here, it is clear that the models have a much poorer fit on both the training and test sets when the coefficient of determination (R^2), Pearson correlation (PCC), and mean-squared errors (MSE) with those in Figure 21.

When we look at Figure 24, we do see that the predicted values of the chatbot SSPA scores are now much more similarly distributed when compared to the predicted SSPA scores on the real conversations. However, we do know that the predicted SSPA scores for the real conversations are not particularly accurate only using this limited feature set restricted to just a few principal components from three feature domains (lexical density, semantic coherence, appropriateness of response); in fact, we previously found that lexical density was not a good predictor for SSPA score performance [22]. It also makes sense that the chatbot predicted scores would be similarly distributed, as the features themselves were similarly distributed, and applying the same model

would lead to a similar distribution of scores. Still, since the fit is poor on the real conversation models themselves, we cannot form any conclusions about whether the chatbot serves as viable alternative to a human clinician until more investigation is done.

6.5 Conclusions

The use of natural language processing and speech processing technologies in clinical work is currently severely limited by the lack of available data on which to train robust and reliable models for the evaluation of mental health. Meanwhile, recent progress with transformer-based large language models has significantly accelerated the pace of research in all areas of natural language processing. With the recent release of the GPT-3 API with fine-tuning capabilities for language generation in specific downstream tasks, we see a new opportunity to augment our real data with artificially generated textual samples with which we can improve our overall language modeling.

However, in order to do so, we must first validate the use of GPT-3 for generating language samples that are clinically relevant. To do so, we took several different approaches, with mixed results. In our digital twin experiments, in which GPT-3 was responsible for generating a full dialogue between two participants, we saw that the AI engine was capable of learning how to imitate the conversation structure and form complete dialogues in most cases; however, we were not able to observe any meaningful difference between conversations that were intended to represent individuals from our three groups (Sz/Sza, BPD, and healthy controls). This was true regardless of what strategy we used to prompt GPT-3.

Therefore, we had to rely on including a human in the loop for our experiments, and

had GPT-3 only play the role of a clinical assessor in the SSPA tasks that are scored in Scenes 2 (new neighbor) and 3 (landlord). Several healthy individuals were recruited in order to have conversations with both versions of these chatbots. We found that the way in which individuals interact by typing on a screen is likely very different than the manner in which they would interact with another person, and the language samples reflect this when we compute our set of features and principal components; this was reflected by the large discrepancies in feature distributions when looking at healthy individuals interacting with a human compared to healthy individuals interacting with the chatbot in many cases. Still, we do see some important similarities, particularly in the areas of lexical density (information packing in utterances), semantic coherence, and appropriateness of response. In most cases, individuals were able to have successful full conversations with the chatbot with few issues.

Still, there is much more investigation to be done in both aspects of this research. For the digital twin experiments, tuning the hyperparameters of the GPT-3 fine-tuning engine could lead to better results for the generation of full conversations, though we were unable to find settings that lead to reasonably good conversation outcomes. Also, as language generating models continue to rapidly improve, we expect to see models that are increasingly capable of understanding the subtle differences in conversation quality between individuals of these different groups. Assuming we are able to successfully generate digital twin conversations, the opportunities to improve the robustness and reliability of language models for clinical applications are nearly endless.

For the chatbot experiments, the collection of more conversation samples will lead to more examples with which the fine-tuning API can learn to respond appropriately to prompts from the user. Similarly, we expect the capabilities of GPT-3 and future

language models to significantly improve chatbot performance. Still, we see that individuals interact quite differently when typing to a screen as opposed to when they converse with a human face-to-face. Assuming that we have a reliably good language generation engine for this application, the next steps would be to create a virtual avatar that listens to the user's spoken prompt (via automatic speech recognition), generates an appropriate response, and replies verbally through use of a text-to-speech engine. We believe this would lead to conversations that better resemble the conversations individuals have with other human beings in a task such as the SSPA evaluation.

While we were not able to show results that definitively demonstrate the capability of GPT-3 to improve clinical language modeling, we have contributed some knowledge about the limitations of this approach while also highlighting much potential. With additional work in this area, we believe it is possible to continue improving language generating models to help augment clinical data sets and build more robust models in real-world settings.

Chapter 7

CONCLUDING REMARKS & FUTURE WORK

7.1 Summary of Research Challenges and Objectives

Due to the incredible burden of severe mental illness and neurodegenerative disease, there is a pressing need for the improvement of healthcare technology and treatment methodologies for a host of debilitating conditions such as schizophrenia, bipolar disorder, dementia, aphasia, and other forms of cognitive impairment. For this reason, there is currently very active research on how to better identify, diagnose, assess, and manage these conditions on a global scale. While trained clinicians are often skilled and capable of managing these conditions, it is important to develop technology and build tools that can aid them in providing better care for impaired individuals and improving their quality of life.

As we identified in Chapter 2, there has been a host of research specifically investigating the potential of automatically computed speech and language metrics for the assessment of cognitive and thought disorders. In recent years, large technological leaps in artificial intelligence, machine learning (particularly deep neural networks) for natural language processing (NLP) and speech signal processing have accelerated the pace of such studies; this has opened up many new opportunities and challenges regarding the implementation and integration of these language technologies into clinical or healthcare workflows.

While many studies effectively demonstrate the strong predictive power of computational speech and language features, to date several researchers have pointed out

that there has been very little in terms of standardizing approaches and obtaining repeatable or robust observations that are made across several studies investigating similar problems [121, 122, 123]. There is also a general interpretability crisis when it comes to machine learning models and deep neural networks, as the inner workings of large AI models are unclear to end users who cannot be certain a model is measuring constructs and making decisions based on truly important factors. For this reason, there has been a lot of work in interpretable or explainable machine learning, but much of it is still in its early stages and insufficient in making state-of-the-art models accessible or useful to users [173]. Another issue we found is that most work in machine learning research for clinical applications is purely data-driven with the primary objective being to classify individuals into a particular diagnostic group; we also find that much of this work does not fully consider if the features measured are of clinical relevance or interpretable. Several additional challenges exist specifically with regard to medical or clinical data [161]; some of these problems include issues with being able to share data due to regulatory restrictions on private health information, little standardization in how data is collected and documented, and a general lack of sufficient data for many clinical populations and comparable data for healthy controls. While some work has been done to generate synthetic data for more robust modeling [163, 162], there is still significant work to be done in addressing many of these challenges.

Our work summarized in this dissertation has attempted to contribute knowledge and bridge some of the gaps that have been identified for making automated language analysis more viable for use in providing objective metrics for clinical applications. Section 7.2 provides a summary of the contributions we have made to these areas.

7.2 Contributions

The work described in the previous chapters of this report has given the details of the main contributions of our work to the body of knowledge in the field of speech and language analysis for the assessment of mental health, as well as the exploration of methods to address the related challenges around real-world implementation and improving model robustness and reliability. Here, we offer a brief summary of these main contributions in the greater context of the field of work.

7.2.1 Measurement Model Framework

The first major aim of this dissertation was the proposed measurement model framework for identifying clinically relevant and interpretable language metrics derived from the theoretical Levelt model of speech production [124]. This is outlined in detail in Chapter 3, in which we introduce the measurement model framework that can ideally be replicated across many studies involving speech and language data samples for clinical applications. Our measurement model framework helps alleviate many of the concerns mentioned above, especially in regard to the question of clinical relevance or interpretability of features. Since many previous studies that examine the same or similar disorders often derive entirely different sets of relevant features with contradictory conclusions, we hope that the data smoothing operation (principal component analysis in our case) over each of the relevant feature domains that were identified will be helpful in reducing the high variance of these feature types. The framework can be adapted to measure different raw features that may be more relevant to a particular disorder or to the nature of the data set that is being studied. We

believe this is a very important step in making automated speech and language analysis more viable for clinical workflows with a standardized approach that with repeatable and interpretable results across many studies.

7.2.2 Holistic Assessment of Mental Health from Language Samples (Upstream and Downstream Problems)

Next, since the vast majority of previous work with speech and language data only focuses on the classification or diagnosis problem, we determined it was important to take a more holistic approach that examines the capability of language analytics to do a much wider range of assessments; namely these are what we called the *upstream* and *downstream* problems. "Upstream" refers to the assessment of neurological changes, symptom severity, or diagnosis of a particular individual given a language data set, while "downstream" refers to the assessment of an individual's capabilities on particular social and functional competency measures. In our work summarized in Chapter 4, we have shown the capability of even a relatively limited language data set to provide insight about all of these different areas of clinical interest.

7.2.3 Real-world Implementation Challenges

We have also noted that there are several real-world implementation challenges for integrating language technology into clinical workflows at scale. One potential pitfall for many language modeling approaches is the requirement to have transcripts of verbal interactions or assessments; due to the cumbersome nature of perfect manual transcription, we are likely to become heavily reliant on automatic speech recognition

(ASR) technology, which is imperfect and noisy by nature. However, large language models are often built on perfectly transcribed text, and large databases of ASR transcriptions are not as easily available on which we can train more robust models. Our work in Chapter 5 outlines some methods we can use to alleviate this problem by simulating ASR-plausible word substitution errors on perfectly transcribed text.

Lastly, regarding the problem of the general unavailability of clinical data and restrictions due to regulatory requirements, we have attempted to use the state-of-the-art in large language modeling (OpenAI's GPT-3 engine) to generate synthetic conversation data that mimics the real samples. While the work here is still quite limited, we have shown some potential ways forward to improve this capability as language models continue to rapidly improve.

7.3 Limitations of Our Work and Recommendations for Future Studies

While this report covers significant progress in the contributions we have made to the field of speech and language analytics, we do not claim to have completely solved all of the problems above. In all of the four aims, there is still significant work to be done to improve the performance capability of language modeling and integration into clinical environments. To date, no form of this technology is widely being used in clinical settings, and we expect that much more needs to be improved in order for it to be truly viable. The work in this area will likely be continually ongoing for the years and decades to come as technological capabilities and our understanding of them improve.

Regarding the measurement model framework for interpretable language metrics (Chapter 3), it is certainly possible to iterate upon the basic outline we have laid out

in our work. This would require greater input from the psychology community and a true standardization of feature types that fall within the domains we identified. The framework we developed worked well and was highly relevant for our study, but a more generalized framework could be more appropriate in a wider range of cases. One major observation we made in our study in Chapter 4 was that the nature of our language data set (the SSPA transcripts) were fairly limited in their utility, since the language was related to a specific-role playing task. More generalized or diverse language data sets (*e.g.* narrative tasks, question answering, word recall, semantic identification, *etc.*) could lead to a slightly different set of important features. The feature selection and smoothing can also be optimized, as it is possible that something may work better or measure more relevant constructs than the PCA approach which we used in our work. Additionally, regarding the Levelt framework for language production, we largely did not consider the articulation domain and related features, as this would require audio recording samples to fully assess; therefore, this whole component of the measurement model framework needs to be developed and validated through further clinical studies.

Implementation challenges continue to be a major hurdle going forward, and we have only briefly touched upon a few of the problems in this area. Our main focus was on the presence of ASR errors leading to noisy transcripts that could impact modeling performance. We covered a method we developed in Chapter 5 to simulate this noise, but the algorithm itself has much room for improvement. The focus was on word substitution errors, but we could also have word insertion, deletion, and boundary errors as well. More sophisticated language modeling could be incorporated into the basic algorithm to generate a wider range of even more plausible ASR errors. Then, the next steps should be to both evaluate the resilience to noise of current state-of-the-art transformer language models, while also considering how to train

models that could improve on this robustness criteria or de-noise corrupted text. One potential avenue is a machine translation approach to de-noising corrupted text.

Lastly, the language generation of synthetic data for clinical applications using GPT-3 is still very limited. We found it very difficult to generate true digital twins of our real conversations, as the subtle linguistic differences between the conversations of our real clinical and control groups were not easily captured by the fine-tuning prompting strategy approaches we used for these experiments. There are potentially many reasons for this; for example, it is likely that the GPT-3 training approach did not have much experience with clinical language data. We expect this to be improved as generative language modeling continues to rapidly evolve and incorporate more text from more sources, but a more in-depth approach focusing on clinical text may be the best approach for improving these models. It is also possible that altering the fine-tuning and prompting strategies may improve the results slightly, though we found this difficult to systematically approach given the limitations of the GPT-3 API. When we moved on to having a human in-the-loop for our chatbot experiments, we ran into a different set of problems; this was mainly due to the fact that human interactions with other humans and a computer screen are drastically different in nature, as expected. One way to improve this is to use a text-to-speech engine and create a virtual avatar that can converse with subjects for a fixed time (as is done in the real SSPA task) and use this virtual assistant for data collection. Finally, we would need to make good use of the additional synthetic data to actually improve our modeling to be more robust.

7.4 Final Remarks and Impact

This dissertation has summarized several different studies that we have undertaken over the last few years to improve the viability of automated speech and language analysis for clinical use cases, particularly in the area of cognitive and thought disorders. While speech and language technology is rapidly evolving with the advents of machine learning and deep learning in recent years, there are still many huge gaps in between technological capability and practical use cases.

As we stated in the beginning, the goal of healthcare is to improve the quality-of-life of individuals afflicted with debilitating conditions. While doctors, nurses, and other healthcare professionals do significant work at all levels to address this problem, the advent of novel non-invasive approaches for automated objective clinical assessments can be a useful tool in improving both diagnostic or assessment capability (upstream of speech) or social and functional competency outcomes (downstream of speech). Therefore, we believe our work has a significant impact in developing the future of digital metrics for improving healthcare outcomes. While there is still significant work to be done, we hope that the measurement model framework, our holistic set of models for patient evaluation, suggestions for improvements to implementation challenges, and synthetic data generation approaches can make this technology more practically viable for use in real-world scenarios, which will eventually lead us to be able to be better equipped to tackle the challenges associated with the burden of mental illness and cognitive thought disorders on society as a whole.

REFERENCES

[1] D. Vigo, G. Thornicroft, and R. Atun, "Estimating the true global burden of mental illness," *The Lancet Psychiatry*, vol. 3, no. 2, pp. 171–178, Feb. 2016.

[2] Center for Behavioral Health Statistics and Quality, "2016 national survey on drug use and health: Methodological summary and definitions," *Substance Abuse and Mental Health Services Administration, Rockville, MD*, 2017.

[3] G. A. Cecchi, V. Gurev, S. J. Heisig, R. Norel, I. Rish, and S. R. Schrecke, "Computing the structure of language for neuropsychiatric evaluation," *IBM Journal of Research and Development*, vol. 61, no. 2/3, pp. 1:1–1:10, Mar. 2017.

[4] H. He, C. Hu, Z. Ren, L. Bai, F. Gao, and J. Lyu, "Trends in the incidence and DALYs of bipolar disorder at global, regional, and national levels: Results from the global burden of Disease Study 2017," *Journal of Psychiatric Research*, vol. 125, pp. 96–105, Jun. 2020.

[5] H. He, Q. Liu, N. Li, L. Guo, F. Gao, L. Bai, F. Gao, and J. Lyu, "Trends in the incidence and DALYs of schizophrenia at the global, regional and national levels: Results from the Global Burden of Disease Study 2017," *Epidemiology and Psychiatric Sciences*, vol. 29, p. e91, 2020.

[6] F. J. Charlson, A. J. Ferrari, D. F. Santomauro, S. Diminic, E. Stockings, J. G. Scott, J. J. McGrath, and H. A. Whiteford, "Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016," *Schizophrenia Bulletin*, vol. 44, no. 6, pp. 1195–1203, Oct. 2018.

[7] T. Vos, A. A. Abajobir, K. H. Abate, C. Abbafati, K. M. Abbas *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016," *The Lancet*, vol. 390, no. 10100, pp. 1211–1259, Sep. 2017.

[8] P. R. Desai, K. A. Lawson, J. C. Barner, and K. L. Rascati, "Estimating the direct and indirect costs for community-dwelling patients with schizophrenia: Schizophrenia-related costs for community-dwellers," *Journal of Pharmaceutical Health Services Research*, vol. 4, no. 4, pp. 187–194, Dec. 2013.

[9] R. Voleti, J. M. Liss, and V. Berisha, "A Review of Automated Speech and Language Features for Assessment of Cognitive and Thought Disorders," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 282–298, Feb. 2020.

[10] B. Elvevåg, P. W. Foltz, D. R. Weinberger, and T. E. Goldberg, "Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia," *Schizophrenia Research*, vol. 93, no. 1-3, pp. 304–316, Jul. 2007.

[11] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.

[12] G. Bedi, F. Carrillo, G. A. Cecchi, D. F. Slezak, M. Sigman, N. B. Mota, S. Ribeiro, D. C. Javitt, M. Copelli, and C. M. Corcoran, "Automated analysis of free speech predicts psychosis onset in high-risk youths," *npj Schizophrenia*, vol. 1, p. 15030, 2015.

[13] C. M. Corcoran, F. Carrillo, D. Fernández-Slezak, G. Bedi, C. Klim, D. C. Javitt, C. E. Bearden, and G. A. Cecchi, "Prediction of psychosis across protocols and risk cohorts using automated language analysis," *World Psychiatry*, vol. 17, no. 1, pp. 67–75, Feb. 2018.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. of 1st International Conference on Learning Representations*, Scottsdale, AZ, USA, 2013, pp. 1–12.

[15] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation." Association for Computational Linguistics, 2014, pp. 1532–1543.

[16] D. Iter, J. Yoon, and D. Jurafsky, "Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia," in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018, pp. 136–146.

[17] R. Voleti, S. Woolridge, J. M. Liss, M. Milanovic, C. R. Bowie, and V. Berisha, "Objective Assessment of Social Skills Using Automated Language Analysis for Identification of Schizophrenia and Bipolar Disorder," in *Proc. Interspeech 2019*, 2019, pp. 1433–1437.

[18] N. Rezaii, E. Walker, and P. Wolff, "A machine learning approach to predicting psychosis using semantic density and latent content analysis," *npj Schizophrenia*, vol. 5, no. 1, p. 9, Dec. 2019.

[19] N. B. Mota, N. A. P. Vasconcelos, N. Lemos, A. C. Pieretti, O. Kinouchi, G. A. Cecchi, M. Copelli, and S. Ribeiro, "Speech Graphs Provide a Quantitative Measure of Thought Disorder in Psychosis," *PLoS ONE*, vol. 7, no. 4, p. e34928, Apr. 2012.

[20] E. S. Kayi, M. Diab, L. Pauselli, M. Compton, and G. Coppersmith, "Predictive Linguistic Features of Schizophrenia," in *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, 2017, pp. 241–250.

[21] T. L. Patterson, S. Moscona, C. L. McKibbin, K. Davidson, and D. V. Jeste, "Social skills performance assessment among older patients with schizophrenia," *Schizophrenia Research*, vol. 48, no. 2-3, pp. 351–360, Mar. 2001.

[22] R. Voleti, S. M. Woolridge, J. M. Liss, M. Milanovic, G. Stegmann, S. Hahn, P. D. Harvey, T. L. Patterson, C. R. Bowie, and V. Berisha, "Language Analytics for Assessment of Mental Health Status and Functional Competency," *Schizophrenia Bulletin*, 2022 Submitted and Under Review.

[23] C. R. Bowie, C. Depp, J. A. McGrath, P. Wolyniec, B. T. Mausbach, M. H. Thornquist, J. Luke, T. L. Patterson, P. D. Harvey, and A. E. Pulver, "Prediction of Real-World Functional Disability in Chronic Mental Disorders: A Comparison of Schizophrenia and Bipolar Disorder," *American Journal of Psychiatry*, vol. 167, no. 9, pp. 1116–1124, Sep. 2010.

[24] R. Voleti, J. M. Liss, and V. Berisha, "Investigating the Effects of Word Substitution Errors on Sentence Embeddings," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 7315–7319.

[25] A. Jeromin and R. Bowser, "Biomarkers in Neurodegenerative Diseases," in *Neurodegenerative Diseases*, P. Beart, M. Robinson, M. Rattray, and N. J. Maragakis, Eds. Cham: Springer International Publishing, 2017, vol. 15, pp. 491–528.

[26] M. Katsuno, K. Sahashi, Y. Iguchi, and A. Hashizume, "Preclinical progression of neurodegenerative diseases," *Nagoya Journal of Medical Science*, vol. 80, no. 3, pp. 289–298, Aug. 2018.

[27] H. H. Dodge, J. Zhu, N. C. Mattek, D. Austin, J. Kornfeld, and J. A. Kaye, "Use of High-Frequency In-Home Monitoring Data May Reduce Sample Sizes Needed in Clinical Trials," *PLOS ONE*, vol. 10, no. 9, p. e0138095, Sep. 2015.

[28] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, Jul. 2015.

[29] W. J. Levelt, "Producing spoken language: A blueprint of the speaker," in *The Neurocognition of Language*, C. M. Brown and P. Hagoort, Eds. Oxford: Oxford University Press, USA, 1999, ch. 4, pp. 83–122.

[30] A. P. Association and A. P. Association, Eds., *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, 5th ed. Washington, D.C: American Psychiatric Association, 2013.

[31] P. T. Trzepacz and R. W. Baker, *The Psychiatric Mental Status Examination*. New York: Oxford University Press, 1993.

[32] V. C. Pangman, J. Sloan, and L. Guse, "An examination of psychometric properties of the Mini-Mental State Examination and the Standardized Mini-Mental State Examination: Implications for clinical practice," *Applied Nursing Research*, vol. 13, no. 4, pp. 209–213, Nov. 2000.

[33] P. S. Mathuranath, P. J. Nestor, G. E. Berrios, W. Rakowicz, and J. R. Hodges, "A brief cognitive test battery to differentiate Alzheimer's disease and frontotemporal dementia," *Neurology*, vol. 55, no. 11, pp. 1613–1620, Dec. 2000.

[34] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, Apr. 2005.

[35] A. M. Kring, R. E. Gur, J. J. Blanchard, W. P. Horan, and S. P. Reise, "The Clinical Assessment Interview for Negative Symptoms (CAINS): Final Development and Validation," *American Journal of Psychiatry*, vol. 170, no. 2, pp. 165–172, Feb. 2013.

[36] B. Kirkpatrick, G. P. Strauss, L. Nguyen, B. A. Fischer, D. G. Daniel, A. Cienfuegos, and S. R. Marder, "The Brief Negative Symptom Scale: Psychometric Properties," *Schizophrenia Bulletin*, vol. 37, no. 2, pp. 300–305, Mar. 2011.

[37] D. Tzur Bitan, A. Grossman Giron, G. Alon, S. Mendlovic, Y. Bloch, and A. Segev, "Attitudes of mental health clinicians toward perceived inaccuracy of a schizophrenia diagnosis in routine clinical practice," *BMC Psychiatry*, vol. 18, no. 1, p. 317, Dec. 2018.

[38] P. D. Harvey and A. Pinkham, "Impaired self-assessment in schizophrenia: Why patients misjudge their cognition and functioning," *Current Psychiatry*, vol. 14, no. 4, pp. 53–59, 2015.

[39] R. Pies, "How "Objective" Are Psychiatric Diagnoses?" *Psychiatry (Edgmont)*, vol. 4, no. 10, pp. 18–22, Oct. 2007.

[40] J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, and B. Schuller, "Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech," *Neurocomputing*, vol. 84, pp. 65–75, May 2012.

[41] D. A. Snowdon, S. J. Kemper, J. A. Mortimer, L. H. Greiner, D. R. Wekstein, and W. R. Markesbery, "Linguistic Ability in Early Life and Cognitive Function and Alzheimer's Disease in Late Life: Findings From the Nun Study," *JAMA*, vol. 275, no. 7, pp. 528–532, Feb. 1996.

[42] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[43] M. A. Covington and J. D. McFall, "Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR)," *Journal of Quantitative Linguistics*, vol. 17, no. 2, pp. 94–100, May 2010.

[44] E. Brunét, *Le Vocabulaire de Jean Giraudoux. Structure et Évolution.* Slatkine, 1978, no. 1.

[45] A. Honoré, "Some Simple Measures of Richness of Vocabulary," *Association for Literary and Linguistic Computing Bulletin*, vol. 7, no. 2, pp. 172–177, 1979.

[46] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of English: The Penn Treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[47] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (DRAFT)*, 3rd ed., Aug. 2017.

[48] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, vol. 1. Edmonton, Canada: Association for Computational Linguistics, 2003, pp. 173–180.

[49] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken Language Derived Measures for Detecting Mild Cognitive Impairment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2081–2090, Sep. 2011.

[50] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic Features Identify Alzheimer's Disease in Narrative Speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, Oct. 2015.

[51] P. Garrard, V. Rentoumi, B. Gesierich, B. Miller, and M. L. Gorno-Tempini, "Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse," *Cortex*, vol. 55, pp. 122–129, Jun. 2014.

[52] M. Asgari, J. Kaye, and H. Dodge, "Predicting mild cognitive impairment from spontaneous spoken utterances," *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, vol. 3, no. 2, pp. 219–228, Jun. 2017.

[53] D. Wechsler, "Wechsler Memory Scale–Third Edition Manual," *San Antonio, TX: The Psychological Corp.*, 1997.

[54] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, Jan. 2000.

[55] K. C. Fraser, J. A. Meltzer, N. L. Graham, C. Leonard, G. Hirst, S. E. Black, and E. Rochon, "Automated classification of primary progressive aphasia subtypes from narrative speech transcripts," *Cortex*, vol. 55, pp. 43–60, Dec. 2012.

[56] V. Berisha, S. Wang, A. LaCross, and J. Liss, "Tracking Discourse Complexity Preceding Alzheimer's Disease Diagnosis: A Case Study Comparing the Press Conferences of Presidents Ronald Reagan and George Herbert Walker Bush," *Journal of Alzheimer's Disease*, vol. 45, no. 3, pp. 959–963, Mar. 2015.

[57] V. Berisha, S. Wang, A. LaCross, J. Liss, and P. Garcia-Filion, "Longitudinal changes in linguistic complexity among professional football players," *Brain and Language*, vol. 169, pp. 57–63, Jun. 2017.

[58] S. Kemper, "Adults' diaries: Changes made to written narratives across the life span," *Discourse Processes*, vol. 13, no. 2, pp. 207–223, Apr. 1990.

[59] S. Kemper and A. Sumner, "The structure of verbal abilities in young and older adults." *Psychology and Aging*, vol. 16, no. 2, pp. 312–322, 2001.

[60] N. E. Carlozzi, N. L. Kirsch, P. A. Kisala, and D. S. Tulsky, "An Examination of the Wechsler Adult Intelligence Scales, Fourth Edition (WAIS-IV) in Individuals with Complicated Mild, Moderate and Severe Traumatic Brain Injury (TBI)," *The Clinical Neuropsychologist*, vol. 29, no. 1, pp. 21–37, Jan. 2015.

[61] V. H. Yngve, "A Model and an Hypothesis for Language Structure," *Proceedings of the American Philosophical Society*, vol. 104, no. 5, 1960.

[62] L. Frazier, *Syntactic Complexity*. Cambridge, U.K.: Cambridge University Press, 1985.

[63] T. Berg, *Structure in Language: A Dynamic Perspective*, 1st ed., ser. Routledge Studies in Linguistics. New York, NY: Routledge, 2009, no. 10.

[64] D. M. Magerman, "Statistical decision-tree models for parsing," in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics -*. Cambridge, Massachusetts: Association for Computational Linguistics, 1995, pp. 276–283.

[65] D. Lin, "On the structural complexity of natural language sentences," in *Proceedings of the 16th Conference on Computational Linguistics -*, vol. 2. Copenhagen, Denmark: Association for Computational Linguistics, 1996, p. 729.

[66] E. Gibson, "Linguistic complexity: Locality of syntactic dependencies," *Cognition*, vol. 68, no. 1, pp. 1–76, Aug. 1998.

[67] N. B. Mota, R. Furtado, P. P. C. Maia, M. Copelli, and S. Ribeiro, "Graph analysis of dream reports is especially informative about psychosis," *Scientific Reports*, vol. 4, no. 1, Jan. 2014.

[68] F. Carrillo, N. Mota, M. Copelli, S. Ribeiro, M. Sigman, G. Cecchi, and D. Fernandez Slezak, "Automated Speech Analysis for Psychosis Evaluation," in *Machine Learning and Interpretation in Neuroimaging*, I. Rish, G. Langs, L. Wehbe, G. Cecchi, K.-m. K. Chang, and B. Murphy, Eds. Cham: Springer International Publishing, 2016, vol. 9444, pp. 31–39.

[69] E. Charniak, "A Maximum-entropy-inspired Parser," in *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, ser. NAACL 2000. Association for Computational Linguistics, 2000, pp. 132–139.

[70] E. Haugen and J. R. Firth, "Papers in linguistics 1934-1951," *Language*, vol. 34, no. 4, pp. 498–502, Oct. 1958.

[71] E. Altszyler, S. Ribeiro, M. Sigman, and D. Fernández Slezak, "The interpretation of dream meaning: Resolving ambiguity using Latent Semantic Analysis in a small corpus of text," *Consciousness and Cognition*, vol. 56, pp. 178–187, Nov. 2017.

[72] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1 (Long Papers). New Orleans, LA, USA: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237.

[73] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers). Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[74] S. Arora, Y. Liang, and T. Ma, "A Simple but Tough-to-Beat Baseline for Sentence Embeddings," in *Proc. of 5th International Conference on Learning Representations*, Toulon, France, 2017, pp. 1–16.

[75] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1 (Long Papers). New Orleans, LA, USA: Association for Computational Linguistics, Mar. 2017, pp. 528–540.

[76] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 670–680.

[77] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco *et al.*, "Universal Sentence Encoder," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 169–174.

[78] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[79] D. Das, N. Schneider, D. Chen, and N. A. Smith, "Probabilistic Frame-semantic Parsing," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 948–956.

[80] A. M. Colman, *A Dictionary of Psychology*. Oxford University Press, 2015.

[81] S. C. Yudofsky, R. E. Hales, and A. P. Publishing, Eds., *The American Psychiatric Publishing Textbook of Neuropsychiatry and Clinical Neurosciences*, 4th ed. Washington, DC: American Psychiatric Pub, 2002.

[82] S. L. Videbeck, *Psychiatric-Mental Health Nursing*. Lippincott Williams & Wilkins, 2010.

[83] M. Mitchell, K. Hollingshead, and G. Coppersmith, "Quantifying the language of schizophrenia in social media," in *Proc.of the 2nd Workshop on Computational*

Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2015, pp. 11–20.

[84] M. Yancheva and F. Rudzicz, "Vector-space topic models for detecting Alzheimer's disease," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 2337–2346.

[85] H. Goodglass and E. Kaplan, "The assessment of aphasia and related disorders," 1983.

[86] L. Hernández-Domínguez, S. Ratté, G. Sierra-Martínez, and A. Roche-Bergua, "Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 10, pp. 260–268, 2018.

[87] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen *et al.*, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 1, pp. 112–124, Mar. 2015.

[88] Y. Tahir, D. Chakraborty, J. Dauwels, N. Thalmann, D. Thalmann, and J. Lee, "Non-verbal speech analysis of interviews with schizophrenic patients," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference On*. IEEE, 2016, pp. 5810–5814.

[89] R. L. Horwitz-Martin, T. F. Quatieri, A. C. Lammert, J. R. Williamson, Y. Yunusova, E. Godoy, D. D. Mehta, and J. R. Green, "Relation of Automatically Extracted Formant Trajectories with Intelligibility Loss and Speaking Rate Decline in Amyotrophic Lateral Sclerosis," in *Proc. Interspeech 2016*. San Francisco, CA, USA: ISCA, 2016, pp. 1205–1209.

[90] S. Sandoval, V. Berisha, R. L. Utianski, J. M. Liss, and A. Spanias, "Automatic assessment of vowel space area," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. EL477–EL483, Nov. 2013.

[91] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," *CUIDADO IST Project Report*, vol. 54, no. 0, pp. 1–25, 2004.

[92] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[93] J. C. Hailstone, G. R. Ridgway, J. W. Bartlett, J. C. Goll, A. H. Buckley, S. J. Crutch, and J. D. Warren, "Voice processing in dementia: A neuropsychological and neuroanatomical analysis," *Brain*, vol. 134, no. 9, pp. 2535–2547, Sep. 2011.

[94] M. N. Stuttle, J. D. Williams, and S. Young, "A framework for dialogue data collection with a simulated ASR channel," in *Eighth International Conference on Spoken Language Processing*. Jeju Island, Korea: ISCA, Oct. 2004, pp. 241–244.

[95] E. Simonnet, S. Ghannay, N. Camelin, and Y. Estève, "Simulating ASR errors for training SLU systems," in *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*. Miyazaki, Japan: European Language Resources Association, May 2018, pp. 3157–3162.

[96] K. López-de-Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay-Torres, P. Martínez-Lage, and H. Eguiraun, "On Automatic Diagnosis of Alzheimer's Disease Based on Spontaneous Speech Analysis and Emotional Temperature," *Cognitive Computation*, vol. 7, no. 1, pp. 44–55, Feb. 2015.

[97] S. V. S. Pakhomov, G. E. Smith, D. Chacon, Y. Feliciano, N. Graff-Radford, R. Caselli, and D. S. Knopman, "Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration," *Cognitive and Behavioral Neurology: Official Journal of the Society for Behavioral and Cognitive Neurology*, vol. 23, no. 3, pp. 165–177, Sep. 2010.

[98] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, G. Szatlóczki, Gréta, E. Biró, F. Zsura, M. Pákáski, and J. Kálmán, "Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech Using ASR," in *Proc. Interspeech 2015*. Dresden, Germany: ISCA, Sep. 2015, pp. 2694–2698.

[99] L. Tóth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatlóczki, Z. Bánréti, M. Pákáski, and J. Kálmán, "A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech," *Current Alzheimer Research*, vol. 15, no. 2, pp. 130–138, 2018.

[100] A. Guidi, N. Vanello, G. Bertschy, C. Gentili, L. Landini, and E. P. Scilingo, "Automatic analysis of speech F0 contour for the characterization of mood changes in bipolar patients," *Biomedical Signal Processing and Control*, vol. 17, pp. 29–37, Mar. 2015.

[101] K. C. Fraser, F. Rudzicz, and E. Rochon, "Using Text and Acoustic Features to Diagnose Progressive Aphasia and its Subtypes," in *Proc. Interspeech 2013*. Lyon, France: ISCA, Aug. 2013, pp. 2177–2181.

[102] R. Vipperla, S. Renals, and J. Frankel, "Longitudinal study of ASR performance on ageing voices," in *Proc. Interspeech 2008*. Brisbane, Australia: ISCA, Sep. 2008, pp. 2550–2553.

[103] V. Young and A. Mihailidis, "Difficulties in Automatic Speech Recognition of Dysarthric Speakers and Implications for Speech-Based Applications Used by the Elderly: A Literature Review," *Assistive Technology*, vol. 22, no. 2, pp. 99–112, 2010.

[104] D. Hakkani-Tür, D. Vergyri, and G. Tur, "Speech-based automated cognitive status assessment," in *Proc. Interspeech 2010*. Makuhari, Chiba, Japan: ISCA, Sep. 2010, pp. 258–261.

[105] L. Zhou, K. C. Fraser, and F. Rudzicz, "Speech Recognition in Alzheimer's Disease and in its Assessment," in *Proc. Interspeech 2016*. San Francisco, CA, USA: ISCA, Sep. 2016, pp. 1948–1952.

[106] B. Mirheidari, D. Blackburn, M. Reuber, T. Walker, and H. Christensen, "Diagnosing People with Dementia Using Automatic Conversation Analysis," in *Proc. Interspeech 2016*, Sep. 2016, pp. 1220–1224.

[107] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Toward the Automation of Diagnostic Conversation Analysis in Patients with Memory Complaints," *Journal of Alzheimer's Disease*, vol. 58, no. 2, pp. 373–387, May 2017.

[108] R. Sadeghian, J. Schaffer, and S. Zahorian, "Speech Processing Approach for Diagnosing Dementia in an Early Stage," *Interspeech 2017*, pp. 2705–2709, Aug. 2017.

[109] J. Weiner, M. Engelbart, and T. Schultz, "Manual and Automatic Transcriptions in Dementia Detection from Speech," in *Proc. Interspeech 2017*. Stockholm, Sweden: ISCA, 2017, pp. 3117–3121.

[110] A. König, N. Linz, J. Tröger, M. Wolters, J. Alexandersson, and P. Robert, "Fully Automatic Speech-Based Analysis of the Semantic Verbal Fluency Task," *Dementia and Geriatric Cognitive Disorders*, vol. 45, no. 3-4, pp. 198–209, 2018.

[111] Y. Jiao, V. Berisha, and J. Liss, "Interpretable phonological features for clinical applications," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 5045–5049.

[112] M. Tu, V. Berisha, and J. Liss, "Interpretable Objective Assessment of Dysarthric Speech Based on Deep Neural Networks," in *Proc. Interspeech 2017*. Stockholm, Sweden: ISCA, Aug. 2017, pp. 1849–1853.

[113] C. M. Corcoran, V. A. Mittal, C. E. Bearden, R. E. Gur, K. Hitczenko, Z. Bilgrami, A. Savic, G. A. Cecchi, and P. Wolff, "Language as a biomarker for psychosis: A natural language processing approach," *Schizophrenia Research*, Jun. 2020.

[114] V. Johansson, "Lexical Diversity and Lexical Density in Speech and Writing: A Developmental Perspective," *Working Papers in Linguistics*, vol. 53, pp. 61–79, 2009.

[115] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing With Compositional Vector Grammars," in *In Proceedings of the ACL Conference*, 2013.

[116] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[117] A. Y. Ng and M. I. Jordan, "On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naïve Bayes," in *Advances in Neural Information Processing Systems*, 2002, pp. 841–848.

[118] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, Nov. 1901.

[119] C.-A. Deledalle, J. Salmon, and A. Dalalyan, "Image denoising with patch based PCA: Local versus global," in *Procedings of the British Machine Vision Conference 2011.* Dundee: British Machine Vision Association, 2011, pp. 25.1–25.10.

[120] T. Takiguchi and Y. Ariki, "PCA-Based speech enhancement for distorted speech recognition." *Journal of multimedia*, vol. 2, no. 5, Sep. 2007.

[121] K. Hitczenko, H. Cowan, V. Mittal, and M. Goldrick, "Automated coherence measures fail to index thought disorder in individuals at risk for psychosis," in *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access.* Online: Association for Computational Linguistics, Jun. 2021, pp. 129–150.

[122] G. M. Stegmann, S. Hahn, J. Liss, J. Shefner, S. B. Rutkove, K. Kawabata, S. Bhandari, K. Shelton, C. J. Duncan, and V. Berisha, "Repeatability of Commonly Used Speech and Language Features for Clinical Applications," *Digital Biomarkers*, vol. 4, no. 3, pp. 109–122, Dec. 2020.

[123] J. Rusz, J. Švihlík, P. Krýže, M. Novotný, and T. Tykalová, "Reproducibility of Voice Analysis with Machine Learning," *Movement Disorders*, vol. 36, no. 5, pp. 1282–1283, May 2021.

[124] W. J. Levelt, "Models of word production," *Trends in Cognitive Sciences*, vol. 3, no. 6, pp. 223–232, Jun. 1999.

[125] G. R. Kuperberg, "Language in Schizophrenia Part 1: An Introduction: Language in Schizophrenia Part 1," *Language and Linguistics Compass*, vol. 4, no. 8, pp. 576–589, Aug. 2010.

[126] L. Weiner, N. Doignon-Camus, G. Bertschy, and A. Giersch, "Thought and language disturbance in bipolar disorder quantified via process-oriented verbal fluency measures," *Scientific Reports*, vol. 9, no. 1, p. 14282, Dec. 2019.

[127] R. E. Hoffman, "A Comparative Study of Manic vs Schizophrenic Speech Disorganization," *Archives of General Psychiatry*, vol. 43, no. 9, p. 831, Sep. 1986.

[128] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[129] N. Dziri, E. Kamalloo, K. Mathewson, and O. Zaiane, "Evaluating Coherence in Dialogue Systems using Entailment," in *Proceedings of the 2019 Conference of the North*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 3806–3812.

[130] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[131] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue *et al.*, "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

[132] E. Merdivan, D. Singh, S. Hanke, J. Kropf, A. Holzinger, and M. Geist, "Human Annotated Dialogues Dataset for Natural Conversational Agents," *Applied Sciences*, vol. 10, no. 3, p. 762, Jan. 2020.

[133] M. Schmidt, *Rey Auditory Verbal Learning Test: A Handbook*. Western Psychological Services Los Angeles, CA, 1996.

[134] R. M. Reitan and D. Wolfson, "The Halstead—Reitan Neuropsychological Test Battery: Research findings and clinical application," in *Specific Learning Disabilities and Difficulties in Children and Adolescents*, 1st ed., A. S. Kaufman and N. L. Kaufman, Eds. Cambridge University Press, Jul. 2001, pp. 309–346.

[135] D. Wechsler, *Wechsler Adult Intelligence Scale: WAIS-IV ; Technical and Interpretive Manual*, 4th ed. San Antonio, Tex. [u.a]: Pearson, 2008.

[136] R. K. Heaton, G. J. Chelune, J. L. Talley, G. G. Kay, and G. Curtiss, *Wisconsin Card Sorting Test Manual: Revised and Expanded.* Odessa, FL: Psychological Assessment Resources Inc, 1993.

[137] E. Strauss, E. M. S. Sherman, O. Spreen, and O. Spreen, *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*, 3rd ed. Oxford ; New York: Oxford University Press, 2006.

[138] B. A. Cornblatt, N. J. Risch, G. Faris, D. Friedman, and L. Erlenmeyer-Kimling, "The continuous performance test, identical pairs version (CPT-IP): I. new findings about sustained attention in normal families," *Psychiatry Research*, vol. 26, no. 2, pp. 223–238, Nov. 1988.

[139] A. J. Snelbaker, G. S. Wilkinson, G. J. Robertson, and J. J. Glutting, "Wide range achievement test 3 (wrat3)," in *Understanding Psychological Assessment.* Springer, 2001, pp. 259–274.

[140] S. R. Kay, A. Fiszbein, and L. A. Opler, "The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia," *Schizophrenia Bulletin*, vol. 13, no. 2, pp. 261–276, Jan. 1987.

[141] L. White, P. D. Harvey, L. Opler, and J. Lindenmayer, "Empirical Assessment of the Factorial Structure of Clinical Symptoms in Schizophrenia," *Psychopathology*, vol. 30, no. 5, pp. 263–274, 1997.

[142] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002.

[143] A. C. Altamura and J. M. Goikolea, "Differential diagnoses and management strategies in patients with schizophrenia and bipolar disorder," *Neuropsychiatric Disease and Treatment*, vol. 4, no. 1, pp. 311–317, Feb. 2008.

[144] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[145] G. Bohouta and V. Këpuska, "Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx)," *Int. Journal of Engineering Research and Application*, vol. 2248–9622, pp. 20–24, Mar. 2017.

[146] C.-H. Li, S.-L. Wu, C.-L. Liu, and H.-y. Lee, "Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 3459–3463.

[147] S. Jung, C. Lee, K. Kim, and G. G. Lee, "An integrated dialog simulation technique for evaluating spoken dialog systems," in *Coling 2008: Proceedings of the Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications*, 2008, pp. 9–16.

[148] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, "From word embeddings to document distances," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. Lille, France: JMLR.org, Jul. 2015, pp. 957–966.

[149] J. Mu, S. Bhat, and P. Viswanath, "Representing Sentences as Low-Rank Subspaces," *arXiv:1704.05358 [cs]*, Apr. 2017.

[150] K. Ethayarajh, "Unsupervised Random Walk Sentence Embeddings: A Strong but Simple Baseline," in *Proceedings of The Third Workshop on Representation Learning for NLP*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 91–100.

[151] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli, "A SICK cure for the evaluation of compositional distributional semantic models," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 216–223.

[152] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 1–14.

[153] N. C. Sanders and S. B. Chin, "Phonological Distance Measures*," *Journal of Quantitative Linguistics*, vol. 16, no. 1, pp. 96–114, Feb. 2009.

[154] B. Allen and M. Becker, "Learning alternations from surface forms with sublexical phonology," *Unpublished manuscript, University of British Columbia and Stony Brook University. Available as lingbuzz/002503*, 2015.

[155] K. C. Hall, B. Allen, M. Fry, S. Mackie, and M. McAuliffe, "Phonological CorpusTools," in *14th Conference for Laboratory Phonology*, Tokyo, Japan, 2015.

[156] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, p. 707, Feb. 1966.

[157] R. L. Weide, "The CMU pronouncing dictionary," *URL: http://www. speech. cs. cmu. edu/cgibin/cmudict*, 1998.

[158] B. Hayes, "Introductory Phonology," *Blackwell Textbooks in Linguistics*, 2009.

[159] Y. Piersmen, "Comparing Sentence Similarity Methods," May 2018.

[160] C. S. Perone, R. Silveira, and T. S. Paula, "Evaluation of sentence embeddings in downstream and linguistic probing tasks," *arXiv:1806.06259 [cs]*, Jun. 2018.

[161] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, "A Review of Challenges and Opportunities in Machine Learning for Health," *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2020, pp. 191–200, 2020.

[162] A. Tucker, Z. Wang, Y. Rotalinti, and P. Myles, "Generating high-fidelity synthetic patient data for assessing machine learning healthcare software," *npj Digital Medicine*, vol. 3, no. 1, p. 147, Dec. 2020.

[163] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, "Synthetic data in machine learning for medicine and healthcare," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 493–497, Jun. 2021.

[164] M. Grieves, "Digital twin: Manufacturing excellence through virtual factory replication," *White paper*, vol. 1, pp. 1–7, 2014.

[165] ——, "Origins of the Digital Twin Concept," 2016.

[166] M. Grieves and J. Vickers, "Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems," in *Transdisciplinary Perspectives on Complex Systems*, F.-J. Kahlen, S. Flumerfelt, and A. Alves, Eds. Cham: Springer International Publishing, 2017, pp. 85–113.

[167] D. Jones, C. Snider, A. Nassehi, J. Yon, and B. Hicks, "Characterising the Digital Twin: A systematic literature review," *CIRP Journal of Manufacturing Science and Technology*, vol. 29, pp. 36–52, May 2020.

[168] B. Björnsson, C. Borrebaeck, N. Elander, T. Gasslander, D. R. Gawel *et al.*, "Digital twins to personalize medicine," *Genome Medicine*, vol. 12, no. 1, p. 4, Dec. 2020.

[169] A. Croatti, M. Gabellini, S. Montagna, and A. Ricci, "On the Integration of Agents and Digital Twins in Healthcare," *Journal of Medical Systems*, vol. 44, no. 9, p. 161, Sep. 2020.

[170] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[171] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[172] F. J. Massey, "Distribution Table for the Deviation Between two Sample Cumulatives," *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 435–441, Sep. 1952.

[173] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2018, pp. 80–89.

APPENDIX A

SAMPLE SSPA TRANSCRIPTS

In this appendix we provide examples of three transcriptions for all three role-playing SSPA exercises. These conversations were previously collected and transcribed at Queen's University (Kingston, ON, Canada) and provided to our group with no identifying information, serving as the basis for our algorithm development and computational analyses. We have one transcript each for a high-performing healthy control, an average-performing individual with BPD, and a low-performing individual with Sz/Sza. Their performance was judged based on their SSPA composite score.

A.1 Transcript examples

A.1.1 Healthy Control

A.1.1.1 Scene 1 (plan activity)

`A:` I'm looking forward to getting together on Friday. What would you like to do?
`S:` Um I think we can go and um have a coffee, yes?
`A:` Um, I did that earlier today maybe we can do something else?
`S:` Um hm, can you suggest something?
`A:` Hmm, don't really have any ideas, right now
`S:` No? Okay. Ugh lets go and have uh something to eat and then we can go to my place and we can watch a movie or we can play a video game or something like that.
`A:` Sure! I'm not very good at video games so maybe a movie.
`S:` Yeah. You can decide or together we can decided uh a movie. Or, if you want to go to the theater we can go there.
`A:` Um yeah lets go for dinner and the theater
`S:` Um yeah that's I think one I like, yeah lets go yeah
`A:` Okay we'll do that... Um what time works for you?
`S:` What time? Ugh I guess I would be good on 7:30

A.1.1.2 Scene 2 (new neighbor)

S: Hi my name is Ika and I guess you are my new neighbour. How are you and what's your name?
A: Hi my names Melissa and yeah I'm your new neighbour.
S: Okay where are you moving from?
A: From Guelph Ontario
S: Oh okay, you're most welcome to this new neighbourhood it's a fantastic neighbourhood. Can I do something for you? Can I offer you coffee or anything else? I'm just cooking dinner if you would like to have dinner, I can cook it for you and give it to you today.
A: Oh wow I really appreciate that that's so nice of you
S: Oh you're welcome ha ha so ugh I think in fifteen or twenty minutes I will be ready with the dinner and I can bring it to you, and if you want to come over to my place you can, no problem. Or I can bring it to your place and we can eat together.
A: Yeah that would be nice, thanks.
S: You're welcome, okay then ah so see you at 8:30, it's around 8 now so see you around 8:30 at my place you can meet with my family too.
A: Thanks, um before you, you, leave I was wondering if you could tell me a little bit about the neighbourhood?
S: Um hm it's a nice neighborhood most of the people live here with families lots of children around so you'll have a little bit of noise too but it's good noise not bad noise. And it's a very safe neighbourhood, uh you can do the um groceries, um, very easily because uh the nearest uh grocery place that Kingston center is just ten,uh minutes away from here you can either walk there or you can go in your car or you can take a bus. It's very easy peasy.
A: Yeah that sounds convenient.
S: Um hm, is there anything else phone or internet you want to know?
A: Um I haven't set up my phone or internet yet.
S: Uh yeah uh at Kingston center there is Bell center you can go there and check with them they have internet as well as uh cell phone connections, so you can talk to them. Or there is another mall, Cat center there are many phone service providers there you can go and, and, have a talk and get the best fee.
A: Yeah that sounds great thanks.
S: You're welcome.
A: I need to head to work the day after tomorrow. Where would I catch the bus?

S: Okay, okay, the bus is right uh outside our building, there is a stop there and you can get bus number two from there, it will take you in twenty five minutes, so no no, yeah.
A: Yeah that...
S: Just one bus yeah ha ha... anything else I can help you with?
A: Um. Is it safe to walk in the neighbourhood?
S: Yes perfectly safe perfectly safe, even uh during uh night sometimes, it is lonely of course but it is perfectly safe. Nothing ever happens there yes. Perfectly safe.
A: And I like to go for runs in the evening.
S: It's okay perfectly safe neighborhood. No problem at all, yup.
A: Okay sounds good... Ah what's the land lord like?
S: He is good.

A.1.1.3 Scene 3 (landlord)

A: This is Mrs. Jones the land lady.
S: Okay. Hi, hi, this Ika calling and I uh just informed you about the leak in my apartment uh two days ago, I guess, and, but nothing has been done so far. And I'm been wonder if, if everything is okay that you're not able to do because usually you respond very well within the same day you are able to get the work done. Is there a problem, if I can help with it, are you sending someone today? Please let me know.
A: Hi um I'm sorry I haven't had enough time to get over and fix it. I've been really busy.
S: Um hm, okay. Are you sending somebody today? Because it is terrible now, it has increased, the leaking has increased terribly so it may cause some damage inside the apartment if you don't send somebody urgently, so if you can't do it just let me know, I can get it done and can give the receipt to you. You can reimburse me.
A: Um so it might take me a week possibly, two weeks before I'll have time to fix it.
S: Oh. But uh what about the damage to the apartment and to my stuff? So are you ready to pay me for that? If it damages my stuff?
A: It didn't sound that bad the last time you called.
S: No but it is uh terrible now. Yeah actually the whole bathroom is sort of flooded with water, sort of. So it can come into my bedroom or the living room, it can damage my things, the carpet, or other things. And uh the carpet is all yours, you know that. So you'll have to pay a lot of money to get it changed or to get it cleaned or whatever it is.
A: That does sound like it's ah quite urgent.
S: Of course it is, and I uh hope you can get it done today. Because it is still morning and you can uh get hold of somebody to do it. Yup.
A: What have you done with the problem so far?
S: I actually, I don't know what to do I just had that uh potty something like that I tried to fix it.But it's not stopping the leakage, and I'm not an expert in this area so I don't know what to do. If you think you can do it, I can ask somebody to uh come and fix it
A: I do all the repairs myself for the whole building.
S: Okay. So do you think you can find a few minutes to come and have a look at it and repair it I can help you with that.
A: Yeah I'm just looking at my schedule now and I could come um tomorrow morning and come and check it out first thing.
S: Okay, tomorrow morning. Okay! So I have informed you if it causes

some damage it's up to you, but okay, tomorrow morning is fine. Yup.
A: Okay. I'll see you at nine am tomorrow first thing.
S: Okay. So I can call my employer and that, that I will be late for work.

A.1.2 Bipolar Disorder

A.1.2.1 Scene 1 (plan activity)

A: Alright, so, I'm looking forward to getting together on Saturday, what do you wanna do?
S: Yeah, someone, what do you wanna do? Ha ha, umm.
A: I don't have any ideas.
S: Hm. (?) there are umm... I can't say any movies I'd want to see, or, um, A: yeah me neither.
S: Umm.. definitely don't want to spend a lot on food.
A: right, well, okay, I'm up for being cheap.
S: Umm... um. Uh, just watch TV. Ha.
A: That sounds good, sounds good.

A.1.2.2 Scene 2 (new neighbor)

S: Hello!
A: Hi!
S: Ha ha ha. Um. Uhh. Definitely not. Ahh, um.
A: Do you live in the building here?
S: Oh, across the street.
A: Oh, across the street, okay.
S: Umm. Are you from... are you from Boniwa, or...?
A: Well not originally, but I've been, uh, living here for about fifteen years now.
S: Hm. Where are you from?
A: Um, Wisconsin.
S: Oh. (?) somebody (?)
A: Not really. It's a nice town.
S: Hm. Uh... it's a long way away from, uh, from Wisconsin.
A: Yeah, well.
S: Ha ha ha. Uh, uh, Wisconsin is further west than (?)

A: Well, it's... it's a ways west.
S: Are you a cheese fan?
A: Um, I like cheese, yes. Ha ha.
S: Ha ha ha. So um.
A: So have you lived here a while?
S: Uhh. Uhh.
A: Can't remember?
S: Ha ha ha. Yeah. Supposedly I'd moved in here that's what I've been told but I don't remember. Ha ha ha. Um... in Baltimore, you know, I was born in Baltimore
A: Oh really?
S: Always planned to get out of Baltimore, never did.
A: Hm.
S: When I look at the weather on TV I'm glad I didn't. um, Baltimore actually is, it's higher on the mountain and, uh or it's a plateau, or uh, um. And the bad weather doesn't get up there so, it's one of the uh, um, one of the safest climates of the rest
A: I never realized that.
S: yeah, um. It's um, a temporal, so yeah, uh, there's summer and we get winter.
A: Yeah.
S: Humidity is bad so... but umm... but um, so it can be more comfortable I swear. But, um, they get a lot of the problems.
A: So we don't get a lot of tornadoes here I guess.
S: What's that?
A: Not, uh, not a lot of tornadoes here?
S: No! um, well not in my area. Ha ha. I don't think so, you know.

A.1.2.3 Scene 3 (landlord)

A: Hello, this is Mrs Jones, the landlady.
S: Hi. Yes, uh, this is Mr Senate, umm... I notified you before, I do have a leak in my ceiling, and, you said you would get back to me but, um, I noticed that, but it still, uh, hasn't been fixed, um, uh, I just want to get an idea of um, uh how long it might be before um, somebody can be out and fix the leak.
A: Yeah, I, uh, apologize Mr Senate that I haven't been by there already, uhm, I'm not really sure how much longer it's gonna take, I have a lot of other repairs ahead of yours right now.
S: Uh. Well, uh, I can understand that, but, um, it is a problem with the leak. Uhm, how, how long do you think it might be, um, after, as far as waiting, as far as whatever you need to get done, as far as other tenants.
A: Um, probably gonna be a couple of weeks.
S: My concern um, is... my concern is that it's water coming from the ceiling, and, um, the leak is getting worse, and I've had it happen before where the, uh, where the ceiling collapsed and it happened, um, I happened to see the crack at the time getting bigger and it collapsed on the bed. uhm, so, i, uh, i do have a concern that it is something that needs to be done in an emergency. um, if it were just a, just a leak, uh, it would be one thing, but i'm really afraid that it's getting worse, that the, uh, weight of the water is going to be a major problem.
A: Okay, well I certainly hope it doesn't turn into that, I mean the last time we talked you sounded like it wasn't that serious.
S: Alright, but um, now, now it's becoming worse, and uh, I am concerned about the danger, and uh, I don't know what damage it might cause too.
A: Right. Well we certainly want to minimize damage. Um, are you seeing any cracks in the ceiling?
S: Yeah. Well, I, uh, like I said, it's getting worse, that's uh, saw a crack yes, and um, so I don't know what might be happening, um, also above as far as, um, behind the ceiling... so that might crack, uh, if, if there's an upstairs neighbour, the upstairs neighbour might end up in my apartment. Ha ha.
A: Yeah, that wouldn't be good. Have you gone up to try to talk to your upstairs neighbour, to see if maybe they have, you know, something running over or something?
S: Well, it, uh, it's been, well it's been regular, so, uhm, it's not, uhm, but I wouldn't feel comfortable talking to them. Um, and, it's not, uh, whether or not they have something that's on, the water might have uh,

accumulated, and there would be no way to see that without getting, uh (?)

A.1.3 Schizophrenia/Schizoaffective Disorder

A.1.3.1 Scene 1 (plan activity)

A: I'm looking forward to getting together Saturday night, what would you like to do?
S: What would I like to do? Uhh, maybe go to uh, a movie?
A: Okay, well, I like the movies as you know, but I've been to the movies twice this week already, is there something else you can think of you might like to do?
S: Uh, maybe go to a restaurant?
A: Okay that sounds great.
S: No, not a restaurant. To uh, to uh Broadway show.
A: Okay. That sounds fun.
S: Yeah so, okay? So, uh, what, what are we doing? We, we, uh,
A: Well, what time should we get together? What time should we meet?
S: We're playing what, as what?
A: Friends.
S: Friends, alright. What were you asking?
A: I was saying what time should we meet.

A.1.3.2 Scene 2 (new neighbor)

S: Well you're welcome to the building.
A: Thank you.
S: Uhh, ha ha. That's it, ha ha that's it.
A: Well, I'm new to the area, can you tell me about the neighbourhood here?
S: Uh, I live here. Uhh, yeah, I just, uh, I just, I live here.
A: Okay.
S: Well welcome, uh, nice to see you.
A: Nice to see you. You know I moved to the area, can you tell me a little bit about the neighbourhood around here?
S: It's, it's just a neighbourhood.

A: Okay. Are there stores nearby, or. . . ?
S: There's, yeah there's stores just like any other uh, just like any other neighbourhood.
A: What kind of stores are nearby?
S: Anything you want. Uh, all the stores, uh, uh, all kinds of stores. I mean, all, all, uh, uh, any, all, uh, all the stores.
A: Okay.
S: Um, anything you, uh uh, anything you want or need.
A: Okay. I'll look for it, and we'll see what I can find.

A.1.3.3 Scene 3 (landlord)

A: Hello this is Mr Jones, the landlord.
S: You talkin' to me?
A: Hello?
S: Yeah hello? Yeah you asked me about, you called me about the leak?
A: Uh you were calling me about the leak, go ahead.
S: I was, I was calling you about the leak?
A: Yes.
S: So, so what, so what are you asking? What are you asking? There's a leak, right?
A: Yes.
S: So, you're asking, uh, you're what are you asking?
A: I'm not asking you anything, you're the one with the leak.
S: No. I have a leak?
A: Yeah.
S: I have a leak?
A: Yeah.
S: And you're, uh, and you're calling me?
A: No, you're calling me.
S: I called you?
A: Hello this is Mr Jones, how can I help you?
S: Yeah I have a leak in my, uh, in my, in my ceiling. Would you, uh, would you be able to come fix it?
A: Um actually I haven't have enough time to come over and fix it because I've been very busy lately.
S: What, uh, uh, when's the next, when could you come?
A: Might be a week, maybe two weeks.
S: What's the earliest you could come?
A: A week or two weeks or more.

S: What's the earliest?
A: That's it, that's the earliest.
S: Uh, uh, would you be able to come, would you be able to come next week?
A: It might be two weeks or more.
S: So, uh, when's the earliest?
A: Like I just said, two weeks or more.
S: Two weeks or more?
A: Mhm.
S: Alright so can you, can I make an appointment?
A: Um, well, I'll just come by when I have time to do it.
S: Umm. Alright. Okay.
A: Are you gonna accept that?
S: When you have time?
A: Okay.

A.2 Feature Principal Components for the Example Transcripts

Table 18: Computed values for the first principal component for each of the seven feature domains for the three example transcripts above. The number of spoken word tokens by the participant is also included.

Participant Type	Words W	Volition PC1	Affect PC1	Lexical Diversity PC1	Lexical Density PC1	Complexity PC1	Semantics PC1	Appropriateness PC1
Control	951	1.11687862	0.539824674	0.611257199	-0.307052564	0.464980159	5.949258518	3.944718272
Bipolar	666	-0.3267335	0.982749702	-0.466725437	0.401973032	-0.662410981	0.311557954	-0.354556849
Sz/Sza	314	-2.0974184	-0.90176629	-4.309847879	-3.379795496	-1.439680111	-0.477851168	1.01868213

In Table 18 we provide the first computed principal component for each feature domain and word token (W) counts for all of the above transcripts. These are the features used in the simplified models from Section 4.1

In Figure 25, we provide a visual representation of the feature domains using a radar chart. As we have scaled the features such that healthy controls have higher values, the larger the area of the bounding shape in the figure, the more control-like a participant is. It is clear from this figure that there are clear differences between the two participants in the clinical group and the healthy control along Volition, Appropriateness of Response, and Semantic Coherence. Comparing the participant

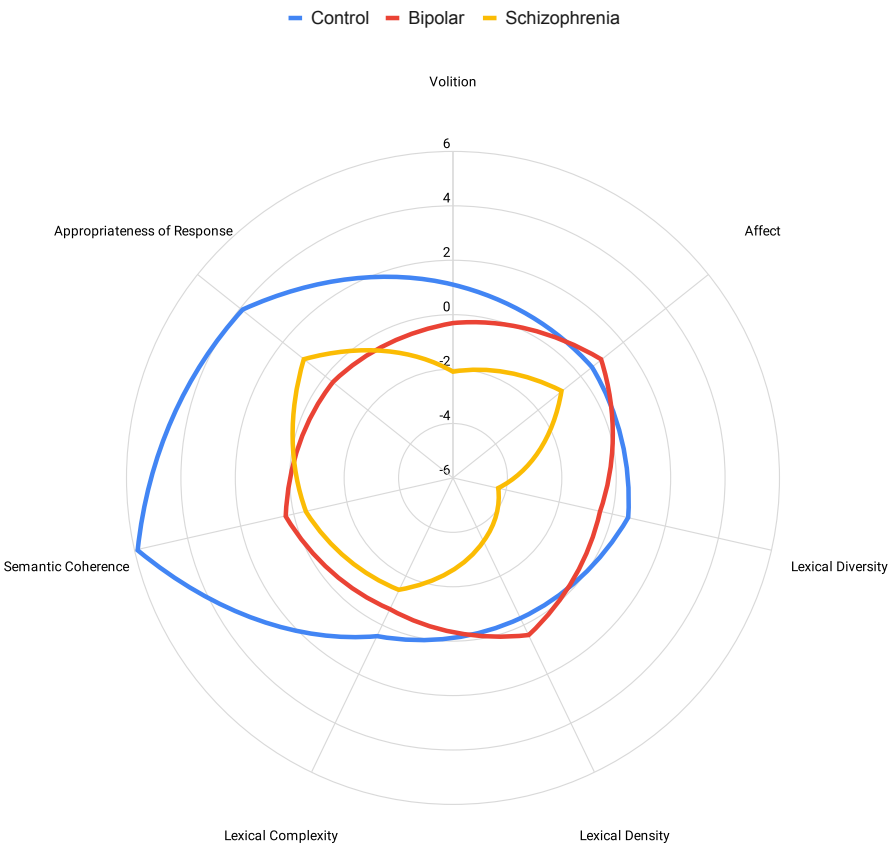


Figure 25: A comparison of the features from the simplified model for the three sample transcripts from Section A.1.

with Sz/Sza and BPD, we see that the individual with BPD is more like the healthy control in Lexical Diversity, Lexical Density, and Affect; the individual with Sz/Sza has lower feature values along all of these dimensions.

BIOGRAPHICAL SKETCH

Rohit Voleti earned his B.S. in Electrical Engineering (2010) and M.S. in Electrical Engineering (2013) from the University of California, Los Angeles. He has been a Ph.D. Student in Electrical Engineering at Arizona State University since Fall 2016 and completed his dissertation with Dr. Visar Berisha and Dr. Julie Liss in the College of Health Solutions (Speech & Hearing Science). Prior to this, he worked a Systems Engineer in the medical device industry in San Diego, CA for Becton Dickinson and as an intern at Advanced Bionics in Valencia, CA.