Supplementary Material: Language Analytics for Assessment of Mental Health Status and Functional Competency

1 Framework for Spoken Language Production

In this work, we make use of a model proposed in [3] 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 [1] and [2], 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 1 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 the paper we focus only on the first two stages, conceptualization and formulation, and leave the representation of the articulation domain for future work.

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.

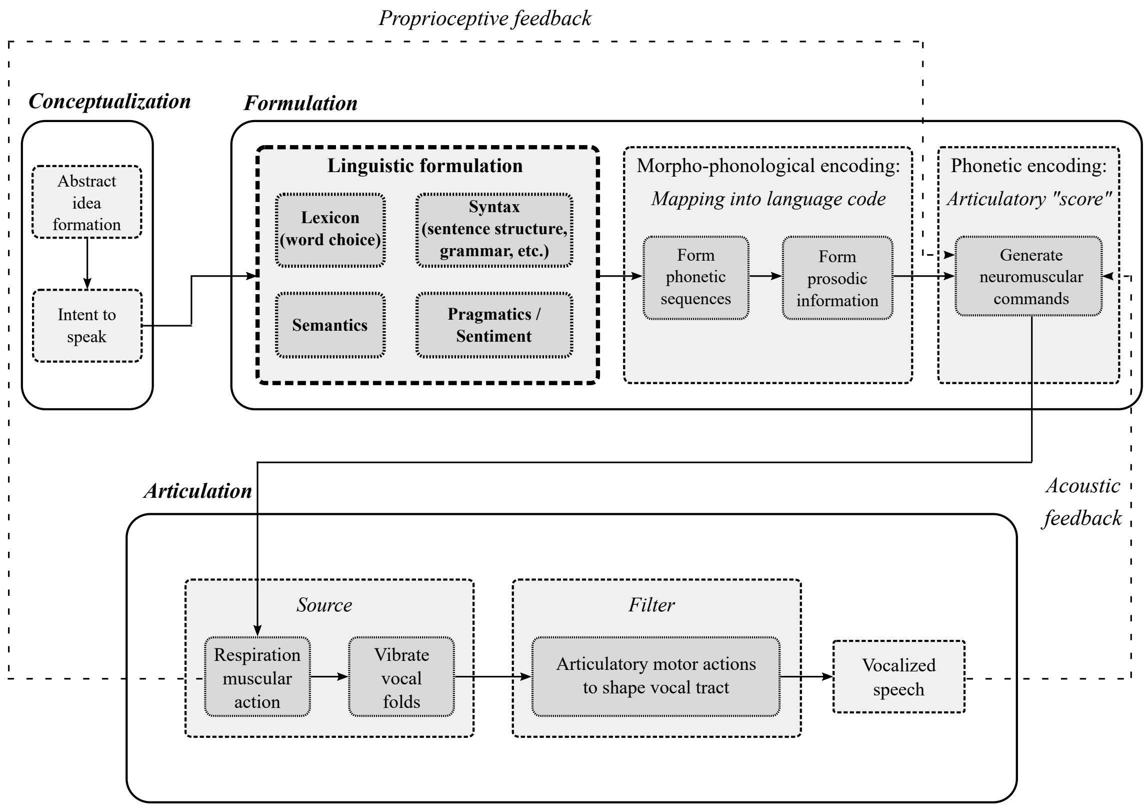


Figure 1: Speech production block diagram model, adapted and modified from [1] and [2]. In our study, we focus on the conceptualization and formulation stages of spoken language production for assessing cognitive health using transcribed conversations.

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 [4]) 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) [5]. This is done in an effort to combat the variability of the lower-level features, as explained in the manuscript.

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.

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. [6]. 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 [7]. 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 1, 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 [6]. Each mood state is associated with a variety of symptoms that impact the speech and language output of that individual [8]. 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 [9]. 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.

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.

2.1 Conceptualization Stage Measurement Model

As discussed in Section 1 and Figure 1, 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.

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) [10, 11]. 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.

Low-level features for affect The Linguistic Inquiry and Word Count (LIWC) tool [12] 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.

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 (1),

$$\mathrm{CosSim}(\mathbf{w}_1, \mathbf{w}_2) = \cos\theta = \frac{\mathbf{w}_1^{\mathrm{T}}\mathbf{w}_2}{\|\mathbf{w}_1\|_2 \, \|\mathbf{w}_2\|_2} \;, \tag{1}$$

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 [10], we considered unweighted averages of the *word2vec* [13], *smooth inverse frequency* (SIF) embeddings [14], and sentence representations generated by the *InferSent* sentence encoder [15].

Additionally, the recent language modeling technique, *Bidirectional Encoder Representations from Transformers* (BERT), proposed in [4], has improved computational performance across a variety of NLP tasks. BERT uses a transformer neural network architecture [16] 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 [17] 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.

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 [18] implementation of BERT from the *transformers* Python library from Huggingface [19]:

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 [20]. 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).

2.2 Formulation State Measurement Model

As discussed in Section 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).

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),defined in Equation (2) as

$$\mathrm{TTR} = \frac{V}{N} \quad . \tag{2}$$

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) [21]: 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) [22]: defined in Equation (3)

  $$\mathrm{BI} = N^{V^{-0.165}} \quad , \tag{3}$$

  in which the exponential reduces the dependence on the total length N. Lower values for BI indicate increased diversity.
- *Honoré's Statistic* (HS) [23]: defined in Equation (4), which emphasizes the use of words that are only spoken once (V_1),

  $$\mathrm{HS} = 100 \log \frac{N}{1 - {V_1}/{V}} \quad . \tag{4}$$

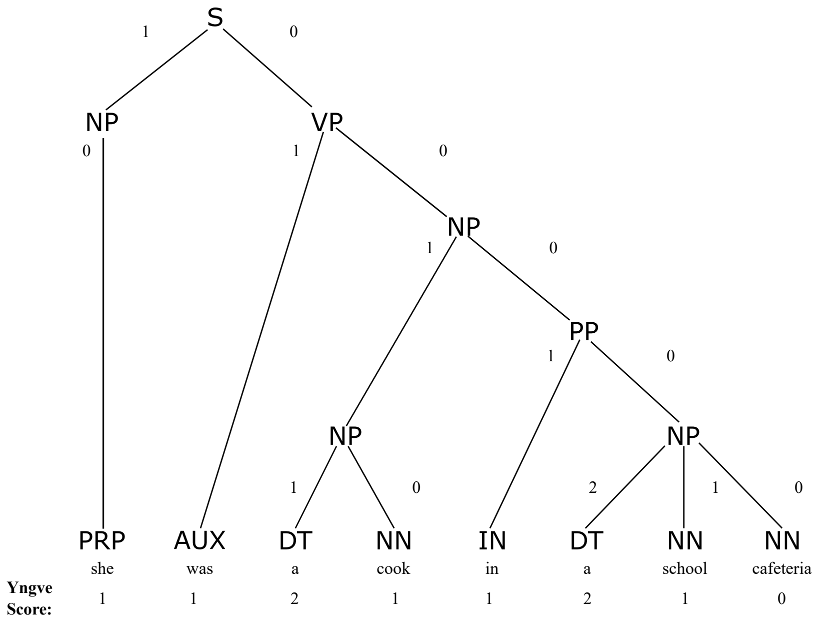


Figure 2: A constituency-based parse tree for a simple sentence. Yngve scoring is a measure of the amount of "left-branching" seen in the parse tree for a given sentence or phrase. Each right-branch is assigned a score of 0, and every consecutive left-branch at a given node has its score incremented by 1. Utterances with more left-branching indicate an increase in embedded clause usage, a measures of linguistic complexity in sentence formulation. The height of the parse tree (3 tiers in this case) is also a useful measure of syntactic complexity.

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 [24] 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 [10].

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 [25]. This allows automatic deconstruction of an utterance into its syntactic structure.

For constituency-based parsing, we use the Stanford parser [25] to decompose each sentence spoken by the participant. Then, Yngve scoring [26] is done for each sentence. An example of constituency-based parsing and Yngve scoring is shown for a simple sentence in Figure 2. 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 [26].

2.3 Feature Composites via Principal Component Analysis

For each of the seven domains, we applied *principal component analysis* (PCA) by domain [5] 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.

3 Model Development

3.1 Data Analysis

Two researchers (Author 1 and Author 5) 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.

3.2 Model Training

3.2.1 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 results are seen for the training (cross-validation) and out-of-sample test sets in the accompanying table for Figure 3 in the paper.

3.2.2 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 for Figure 4 of the paper show the ROC AUC for the two models on the out-of-sample data along with weighted precision and recall, and $F1$ score

3.2.3 Final Models

For reference, we list the final models that were developed from the training samples provided to Author 5.

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):

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

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

Neurocognitive Composite Score Prediction:

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

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:

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

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

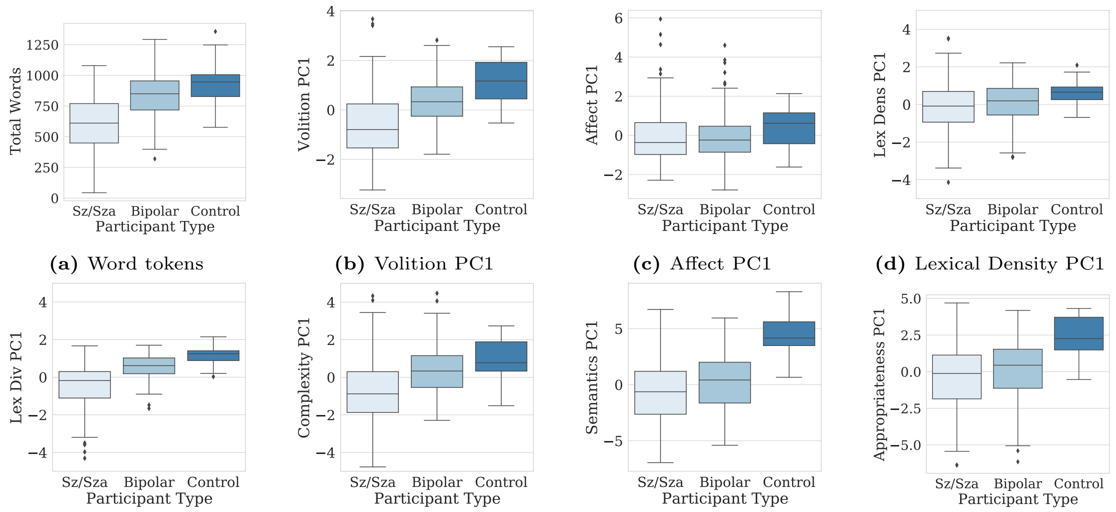
$$\text{SLOF Interpersonal} = 2.09122 + 0.08016 * \sqrt{W} + 0.10498 * \text{Semantics.PC2}$$

SLOF - Work Skills Subscale

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

4 Interpretability Analysis

With a focus on interpretability, we performed an additional analysis in which we asked the algorithm designer to use the same training set to develop models using only the first principal component from each feature domain. By using only a single component for each of the feature domains, we can evaluate how each feature domain impacts predictive outcomes in these simplified models. A set of



(a) Word tokens **(b)** Volition PC1 **(c)** Affect PC1 **(d)** Lexical Density PC1

(e) Lexical diversity PC1 **(f)** Complexity PC1 **(g)** Semantics PC1 **(h)** Appropriateness PC1

Figure 3: Distributions of the number of word tokens (W) and of the simplified 1-dimensional composite features associated with each of the seven feature domains used in our model development.

box plots is shown in Figure 3 for the distribution of each of the first principal components for the seven feature domains plus the distributions for W across all three groups. These components are then used to develop and analyze the simplified models discussed below.

Table 1 shows the results comparing the fit of the original models to the simplified models developed for this analysis. In most cases, we see that the models perform similarly, but often with a poorer fit than the complete models used in the main paper. However, these models are easier to evaluate for interpretability as the composites for the domains in the Levelt framework are 1-dimensional.

4.1 Interpretation of the Simplified Models

Upstream Models

Classification Model 1 (Clinical vs. Control):

$$\begin{aligned}\text{logit} =& 8.709480 - 0.005031 * W - 0.71277 * \text{ Appropriateness.PC1}- \\ & 0.332015 * \text{Semantics.PC1}\end{aligned}$$

$$\text{class} = \begin{cases} \text{control} & \text{logit} < 3.2320915 \\ \text{clinical} & \text{logit} \geq 3.2320915 \end{cases}$$

Interpretation: We see that increased values for W, the Appropriateness PC1, and the Semantic PC1 variables all result in a lower value of the logit for this model, meaning that participants that scored higher in these variables are more likely to be classified as a healthy control. This matches the distributions observed in Figures 3a, 3g, and 3h; it is also consistent with previous literature on the impact of Schizophrenia and Bipolar Disorder on language.

Table 1: The results for the original and simplified models defined in Sections 3.2.3 and 4.1. For models that did not change, the Pearson correlation coefficient (PCC) and mean squared error (MSE) results are represented with dashes.

(a) Downstream regression results

		PCC		**MSE**	
		Orig.	*Simp.*	*Orig.*	*Simp.*
SSPA Avg.	Train	0.787	0.745	0.178	0.258
	Test	0.785	0.791	0.171	0.206
SLOF:					
Interpersonal	Train	0.473	0.445	0.511	0.528
	Test	0.569	0.524	0.493	0.524
Activites	Train	0.647	-	0.160	-
	Test	0.572	-	0.211	-
Work	Train	0.535	0.449	0.734	0.826
	Test	0.352	0.308	0.830	0.893
Fx Composite	Train	0.608	0.612	2.507	2.487
	Test	0.616	0.571	2.422	2.584

(b) Upstream regression results

		PCC		**MSE**	
		Orig.	*Simp.*	*Orig.*	*Simp.*
Neurocog. Composite	Train	0.621	0.555	0.623	0.702
	Test	0.674	0.678	0.682	0.723
PANSS:					
Pos. Symptoms Mean	Train	0.497	0.515	0.580	0.550
	Test	0.509	0.492	0.559	0.549
Neg. Symptoms Mean	Train	0.718	-	0.487	-
	Test	0.767	-	0.476	-

(c) Classification model results

		Precision		**Recall**		**F1**		**AUC**	
		Orig.	*Simp.*	*Orig.*	*Simp.*	*Orig.*	*Simp.*	*Orig.*	*Simp.*
Clinical v. Control	Train	0.971	0.951	0.971	0.797	0.971	0.851	0.856	0.850
	Test	0.968	0.913	0.969	0.802	0.968	0.834	0.903	0.849
BD v. Sz/Sza	Train	0.730	0.740	0.730	0.740	0.729	0.740	0.730	0.740
	Test	0.672	0.687	0.671	0.682	0.670	0.680	0.670	0.681

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

$$\text{logit} = 3.936189 - 0.005431 * W - 0.284251 * \text{Appropriateness.PC1}$$

$$\text{class} = \begin{cases} \text{BPD} & \text{logit} < -0.151383015 \\ \text{Sz/Sza} & \text{logit} \geq -0.151383015 \end{cases}$$

Interpretation: Increased values for W and Appropriateness PC1 again result in a lower logit value, indicating that more appropriate responses and increased verbosity indicate a greater likelihood of a participant being classified as part of the Bipolar group, matching the distributions of these variables observed in Figures 3a and 3h, where the Bipolar group was closer to the Healthy group.

Neurocognitive Composite Score Prediction:

$$\text{Cog Comp} = -2.4924971 + 0.0024341 * W$$

Interpretation: The simplified version of the cognitive composite score prediction depends only on verbal output (W), and we see that increased verbosity results in a higher composite score. We see that those with BD have a higher cognitive composite score on average than those with Sz/Sza (from Table 1 in the main paper), and we would expect healthy controls to score even higher on this assessment, which agrees with the distributions we see in Figure 3a. In the case of this model, imposing the constraint that each domain is modeled as with a one dimensional principal component results in a simplistic model that does not include any of the measured domains. That is, whatever information exists in the language sample about the neurocognitive composite, it is not captured in the principal component of each domain. This provides additional evidence that this is likely not the ideal task for assessing cognitive status.

PANSS Positive Symptoms Mean Prediction:

$$\text{PosMean} = 3.56790 - 0.06353 * \sqrt{W} - 0.08898 * \text{Appropriateness.PC1}$$

Interpretation: We see that an increased Appropriateness PC1 value and increased W both lead to a lower overall rating for positive symptom severity. Healthy controls are expected to have the most appropriate responses and highest verbal output and are also expected to have the lowest severity in terms of symptom ratings. This is consistent with the model.

PANSS Negative Symptoms Mean Prediction:

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

Interpretation: The simplified model does not differ from the original model in this case. Neither used any of the principal components from our feature domains and depend solely on $\sqrt{W}$. Again we see that increased verbosity would predict a lower negative symptom severity, which is in line with what we expect from our BD and Sz/Sza groups.

Downstream Models

Average SSPA Score:

$$\text{SSPA Avg. Score} = 0.15 * \sqrt{W} + 0.085640 * \text{Appropriateness.PC1}$$

Interpretation: The SSPA score is predicted with only the $\sqrt{W}$ and Appropriateness PC1 variables in the simplified model. As expected, those with more appropriate responses and a higher verbal output scored better on this task, as reflected by this model.

SLOF - Fx Composite and all Subscales

$$\text{SLOF Fx} = 6.51938 + 0.23913 * \sqrt{W} + 0.15281 * \text{Appropriateness.PC1}$$
$$\text{SLOF Activities} = 2.82305 + 0.06852 * \sqrt{W} + 0.03072 * \text{Appropriateness.PC1}$$
$$\text{SLOF Interpersonal} = 2.53423 + 0.06374 * \sqrt{W} + 0.04371 * \text{Semantics.PC1}$$
$$\text{SLOF Work} = 1.19661 + 0.10569 * \sqrt{W} + 0.08122 * \text{Appropriateness.PC1}$$

Interpretation: All of the models predicting the SLOF subscales and functional composite score follow a similar pattern to the previously discussed prediction models. We see that an increased verbosity and appropriateness of response (indicated by $\sqrt{W}$ and Appropriateness PC1) predict better SLOF outcomes for the composite score and two of the subscales (activities and work tasks). Additionally, greater semantic coherence (indicated by Semantics PC1) also leads to better outcomes for the Interpersonal relations subscale. This is in line with the distributions for W, Semantics PC1, and Appropriateness PC1 we see in the box plots in Figures 3a, 3g, and 3h.

4.2 Example Transcripts

In this section we provide examples of three transcriptions for all three role-playing SSPA exercises. 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.

4.2.1 Healthy Control

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

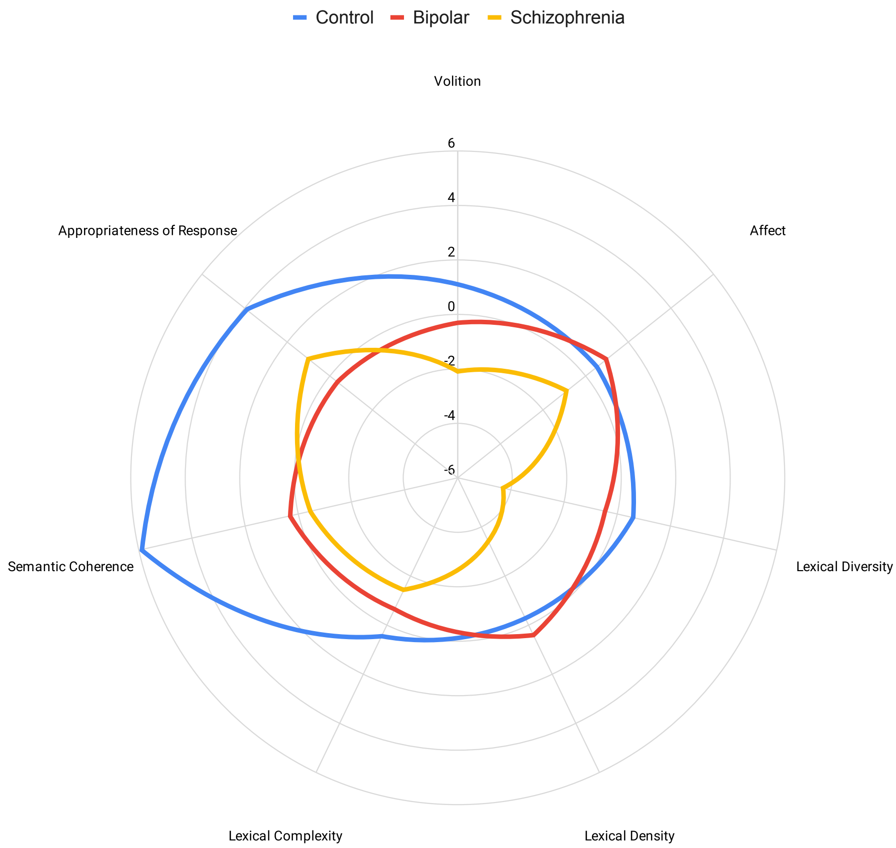


Figure 4: A comparison of the features from the simplified model for the three sample transcripts from Section 4.2.

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.

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.

4.2.2 Bipolar Disorder

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.

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.

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 (?)

4.2.3 Schizophrenia/Schizoaffective Disorder

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.

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.

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.

4.3 Feature Principal Components for the Example Transcripts

Table 2: 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 2 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 4, 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 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.

References

[1] Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. *Speech Communication.* 2015;71:10–49.

[2] Voleti R, Liss JM, Berisha V. A Review of Automated Speech and Language Features for Assessment of Cognitive and Thought Disorders. *IEEE Journal of Selected Topics in Signal Processing.* 2020;14:282–298.

[3] Levelt WJ. Models of word production. *Trends in Cognitive Sciences.* 1999;3:223–232.

[4] Devlin J, Chang MW, Lee K, Toutanova K. 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*;1 (Long and Short Papers)(Minneapolis, MN, USA):4171–4186Association for Computational Linguistics 2019.

[5] Pearson K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science.* 1901;2:559–572.

[6] Association AP, Association AP. , eds.*Diagnostic and Statistical Manual of Mental Disorders: DSM-5.* Washington, D.C: American Psychiatric Association5th ed ed. 2013.

[7] Kuperberg GR. Language in Schizophrenia Part 1: An Introduction: Language in Schizophrenia Part 1. *Language and Linguistics Compass.* 2010;4:576–589.

[8] Weiner L, Doignon-Camus N, Bertschy G, Giersch A. Thought and language disturbance in bipolar disorder quantified via process-oriented verbal fluency measures. *Scientific Reports.* 2019;9:14282.

[9] Hoffman RE. A Comparative Study of Manic vs Schizophrenic Speech Disorganization. *Archives of General Psychiatry.* 1986;43:831.

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

[11] Fraser KC, Meltzer JA, Rudzicz F. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease.* 2015;49:407–422.

[12] Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology.* 2010;29:24–54.

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

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

[15] Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. 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):670–680Association for Computational Linguistics 2017.

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. in *Advances in Neural Information Processing Systems*:5998–6008 2017.

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

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

[19] Wolf T, Debut L, Sanh V, et al. HuggingFace's transformers: State-of-the-art natural language processing. *ArXiv.* 2019;abs/1910.03771.

[20] Merdivan E, Singh D, Hanke S, Kropf J, Holzinger A, Geist M. Human Annotated Dialogues Dataset for Natural Conversational Agents. *Applied Sciences.* 2020;10:762.

[21] Covington MA, McFall JD. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics.* 2010;17:94–100.

[22] Brunét E. *Le Vocabulaire de Jean Giraudoux. Structure et Évolution.* No. 1Slatkine 1978.

[23] Honoré A. Some Simple Measures of Richness of Vocabulary. *Association for Literary and Linguistic Computing Bulletin.* 1979;7:172–177.

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

[25] Socher R, Bauer J, Manning CD, Ng AY. Parsing With Compositional Vector Grammars. in *In Proceedings of the ACL Conference* 2013.

[26] Yngve VH. A Model and an Hypothesis for Language Structure. *Proceedings of the American Philosophical Society.* 1960;104.