Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature: DocuSigned by: Linying Li 11A0002129114EC...

Linying Li
Name

6/22/2023 | 4:45 PM EDT
Date

Title Language as a Subtyping Tool and a Potential Predictor of Treatment Outcome in Depression: Using Large Language Models to Harvest the Predictive Power of Language

Author Linying Li

Degree Master of Arts

Program Psychology

Approved by the Committee

DocuSigned by:
Phillip Wolff
A906AD9BE70D49C...

Phillip Wolff
Advisor

DocuSigned by:
W. Edward Craighead
638B58A98CB64BD...

W. Edward Craighead
Committee Member

DocuSigned by:
Lynne Nygaard
E89C1224B387420...

Lynne Nygaard
Committee Member

Committee Member

Committee Member

Committee Member

Accepted by the Laney Graduate School:

_____

Kimberly Jacob Arriola, Ph.D, MPH
Dean, James T. Laney Graduate School

_____

Date

Language as a Subtyping Tool and a Potential Predictor of Treatment Outcome in Depression: Using Large Language Models to Harvest the Predictive Power of Language

By

Linying Li

B.A., Reed College, 2020

Committee:

Phillip Wolff, PhD (Advisor)

W. Edward Craighead, PhD, ABPP

Lynne Nygaard, PhD

An abstract of

A thesis submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Master of Arts

in Psychology

2023

Abstract

Language as a Subtyping Tool and a Potential Predictor of Treatment Outcome in Depression: Using Large Language Models to Harvest the Predictive Power of Language
By Linying Li

Major depressive disorder (MDD) is a highly debilitating condition. Early treatment optimization is crucial for a favorable prognosis, but reliably predicting who is most likely to benefit from which treatment remains a major challenge. One way to address the problem is through a better understanding of the heterogeneity in the disease. Previous research identified language use as a potential indicator of individual differences in depression, and recent technological advancements permit a more systematic approach to the use of language in this regard. In the current study, we demonstrate how large language models (LLMs) can be used to identify sub-types of depression in the early stages of treatment based on people's natural speech productions. We introduce a computational technique for determining the relative similarity of two narratives by measuring how one narrative affects an LLM's ability to predict sentences in another narrative when it is used as a context. The resulting narrative similarities were analyzed using hierarchical clustering to reveal three major subgroups of depression. Subsequent feature analyses indicated distinguishing semantic and syntactic properties of each cluster and predictions about future remission status. The findings demonstrate how AI models applied to the analysis of people's natural speech can be used in subtyping and predicting treatment outcomes for depression.

Keywords: depression, disease heterogeneity, outcome prognosis, language, large language models

Language as a Subtyping Tool and a Potential Predictor of Treatment Outcome in Depression: Using Large Language Models to Harvest the Predictive Power of Language

By

Linying Li

B.A., Reed College, 2020

Committee:

Phillip Wolff, PhD (Advisor)

W. Edward Craighead, PhD, ABPP

Lynne Nygaard, PhD

A thesis submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Master of Arts

in Psychology

2023

Table of Contents

List of Figures

List of Tables

Language as a Subtyping Tool and a Potential Predictor of Treatment Outcome in Depression: Using Large Language Models to Harvest the Predictive Power of Language

Introduction

Overview

Major depressive disorder (MDD) is the leading cause of disability (Friedrich, 2017). Those suffering from MDD experience depressed moods and loss of interest for prolonged periods of time. During depressive episodes, many have to deal with extreme difficulties in various essential aspects of life, including but not limited to maintaining healthy interpersonal relationships and meeting expectations at work or in educational settings. When looking at disability-adjusted life years (DALY), a summary health indicator evaluating overall disease burden (Murray & World Health Organization, 2002), MDD was associated with the highest number of DALYs among various mental and addictive disorders (Rehm & Shield, 2019). It is also associated with shorter life expectancy (Laursen et al., 2016). In addition to its debilitating nature, the increase in prevalence of the disorder makes the situation even more pressing. A recent meta-analysis surveying publications between 1937 and 2018 showed a significant increase in the likelihood of experiencing depression (Moreno-Agostino et al., 2021). Furthermore, cases of depression have shown pronounced increases since the start of the COVID-19 pandemic (Bueno-Notivol et al., 2021). Given the gravity and prevalence of the disease, the need for adequate preventive measures and treatment plans in response to the current situation of MDD has become increasingly urgent.

Over the years, multiple methods to treat MDD have been established (e.g., psychotherapy, antidepressants, electroconvulsive therapy), and research suggests that personalized optimization of the treatment plan at an early stage is crucial for ensuring treatment efficacy in depression and a failure to do so could lead to mistreatment (Habert et al., 2016; Kraus et al., 2019; Paris, 2014). However, reliably predicting who is most likely to benefit from treatment remains an unresolved task. Ensuring treatment efficacy at the individual level remains a major challenge (Cuijpers et al., 2020; Rost et al., 2023). To better understand the condition and to improve treatment optimization, researchers have long been trying

to profile the heterogeneity within depression (e.g., ten Have et al., 2016; Vicent-Gil et al., 2020), but the existing attempts have limited utility in predicting treatment outcome (e.g., Groves et al., 2018; Uher et al., 2012). Moving forward, researchers recommended exploring novel data types and incorporating various types of predictors in the prediction process (Rost et al., 2023).

In the quest for a better understanding of depression, we could turn to another potential indicator of depression: language. Existing literature demonstrated discriminative capabilities of language features in distinguishing individuals with depression from other groups (e.g., Mariani et al., 2020; Smirnova et al., 2018) and differentiating depression severity (e.g., Pulverman et al., 2015; Rude et al., 2004). These successes suggest that language features may be used to detect the heterogeneity within depression. However, language has not been studied systematically in this regard, likely due to limitations in previously existing methodology. Opportunely, recent advancements in language technology have made it possible for methodological shifts, allowing us to move beyond the simple keyword lists to examine patterns in linguistic behaviors more precisely and at a much greater scale than what could be achieved before (Johns et al., 2020). These new approaches permit a more systematic approach to the use of language as a predictive indicator of depression.

Given the pressing nature of the problem and recent technological advancements, the current study explores the use of language in addressing the challenges of understanding and treating depression through the application of new computational methods. Importantly, considering the still-existing limitations and the fast-evolving nature of language technology, it is not the goal of the study to derive a definitive conclusion about the utility of language in deepening our understanding of depression; rather, we hope to develop a method that will allow us to efficiently apply future language models and speech samples as they become available.

Background

In the sections below, I will first discuss the evidence for the heterogeneous nature of depression. Then, I will summarize the efforts made to translate research on heterogeneity within depression to improve treatment efficacy. Thirdly, I will discuss studies examining language use in depressed populations and provide a case for exploring language as a potential subtyping tool and a predictor of treatment outcome. Stemming from the research on language and depression, I will discuss the importance of attending to context while studying language use. Finally, I will give an overview of the methodological shifts in the study of language that are relevant to the current project.

Evidence for heterogeneity in depression. According to the fifth edition of *Diagnostic and statistical manual of mental disorders* (DSM-5), one meets symptom criteria for MDD if five or more of the following symptoms (which must include one of the two asterisked symptoms) are present: depressed mood*, loss of interest in pleasurable activities*, significant weight or appetite changes, sleep disturbance, significant psychomotor agitation or retardation, low energy, a sense of worthlessness or guilt, decreased ability to concentrate, and suicidal thoughts (American Psychiatric Association, 2013). In theory, there are 227 possible ways to meet symptom criteria for MDD, and it is possible for two individuals who share no symptoms to be both diagnosed with MDD. Though certain symptom combinations are infrequently (or even never) observed clinically (Zimmerman et al., 2015), there still exists a significant number of ways to meet the MDD symptom criteria, alluding to the heterogeneity of the condition.

In addition to the polythetic nature of how MDD is defined, research has provided empirical evidence for the heterogeneous nature of depression (Harald & Gordon, 2012; Quinn et al., 2014). Many attempts have been made to characterize the heterogeneity of depression in terms of subtypes (e.g., ten Have et al., 2016; Wadsworth et al., 2001). Early attempts to do so were often symptom-based, describing depression subtypes in terms of symptom severity, typicality and/or presence of comorbidity (Lamers et al., 2012; Rodgers et al., 2014; ten Have et al., 2016; Wadsworth et al., 2001). More recently, researchers

have investigated possible neurocognitive indicators of subtypes (Baller et al., 2021; Vicent-Gil et al., 2020). Baller and colleagues (2021) identified three subtypes of depression based on participants' accuracy and speed in an fMRI n-back working memory task. Vicent-Gil and colleagues (2020) conducted a cluster analysis which revealed subgroups of depression, characterized by cognitive functioning and treatment resistance. While the existence of specific subtypes of depression remains unresolved (Beijers et al., 2019; van Loo et al., 2012), the heterogeneous nature of depression is widely acknowledged.

Translating knowledge about heterogeneity to treatment optimization. A better understanding of the individual differences in MDD could lead to improvements in treatment efficacy. The goal to improve and promote personalized medicine is in line with the goal of Research Domain Criteria (RDoC; Insel et al., 2010), a research framework launched by the National Institute of Mental Health (NIMH) to encourage the use of integrative methods to better understand mental disorders. As early treatment optimization results in faster relief from depression (Habert et al., 2016; Kraus et al., 2019), translating knowledge about heterogeneity is very much needed.

In a recent review, Rost and colleagues (2023) detailed current approaches to define and predict treatment outcomes in MDD. Though Rost et al. (2023) restricted the review to only one of the common treatments to depression – pharmacological treatments – the issues they raised regarding the current inability to reliably predict treatment outcomes are still relevant when considering other treatments. In short, reliably predicting who is most likely to benefit from which treatment remains an unresolved task, and Rost and colleagues identified the main challenge to be the lack of effective translational efforts. In response to some of the challenges they identified, Rost and colleagues recommended exploring new indicators of disease in predictive modeling. Indeed, the predictors of depression most often investigated--symptom presentation, genotype, neuroimaging data, clinical measures--only explain a limited amount of variance in treatment outcomes (e.g., Gao et al., 2018; Groves et al., 2018; Uher et al., 2012). Identifying

and evaluating novel predictors would be a meaningful next step to improving outcome prognosis in depression.

Language use in depressed populations and its potential in clinical applications. Researchers have identified several language features associated with depressed speech. A meta-analysis surveying 21 studies and a total of 3758 participants identified a positive correlation between first-person singular pronoun use and depression (Edwards & Holtzman, 2017). The results support the notion that those who are depressed tend to be self-focused (Greenberg & Pyszczynski, 1986; Watkins & Teasdale, 2004). Increased use of negative emotion words relative to positive words – or negative bias – has also been observed in the writing samples and speech of individuals who are depressed and at risk for suicide (Baddeley et al., 2011; Kauschke et al., 2018; Kim et al., 2019). In addition to negative bias, researchers have noted that individuals who are depressed tend to use atypical word order, increase their use of the past tense, and use atypical sentence structure (Smirnova et al., 2018, 2019; Trifu et al., 2017). The characteristics mentioned above were found to be consistently associated with depressed affect, both acute and chronic (Newell et al., 2018), showing that depression has a robust influence on one's linguistic expression.

Language features can not only differentiate between depressed and healthy populations, they can also be used to differentiate types of depressive moods. For example, linguistic features can be used to distinguish individuals who are depressed from individuals who are merely having a temporary negative mood (Bernard et al., 2016; Smirnova et al., 2018). Linguistic features can also differentiate individuals with unipolar depression from bipolar disorder based on references to bodily activities and sensations (Mariani et al., 2020). Language features have also been used to differentiate individuals with depression from individuals with anxiety by focusing on sadness-related words (Sonnenschein et al., 2018). While individual language features can often point to depression, they do so best in combination with other language features. For example, first-person pronouns are not able on their own to distinguish depression

from other mental disorders (Lyons et al., 2018), suggesting that selecting what language features to include is crucial to the success of the classification task.

Beyond their discriminative capabilities, language features can be used to measure symptom severity, as demonstrated by Rude and colleagues (2004), who examined written essays of college students who were depressed, formerly depressed, and never depressed at the time of the study. They found that the use of negatively valenced words by the formerly-depressed group did not differ from the never-depressed group, whereas a significant difference was observed between the currently-depressed and the never-depressed, suggesting that linguistic markers of depression are related to current status of depression. However, the use of first-person pronouns among the formerly-depressed group was still elevated, which is probably associated with an ongoing self-focus tendency and increased risk for depression.

The fact that linguistic expression is sensitive to depression severity suggests that it might be used to monitor treatment progress. In a sample of childhood sexual abuse survivors, Pulverman and colleagues (2015) found that a decrease in first-person pronoun use and an increase in positive emotion words indicated decreased depression symptoms and that a reduction of negative emotion words was associated with alleviated sexual dysfunction. Relatedly, Demiray & Gençöz (2018) analyzed clients' speech during psychotherapy and found that by the 15th session, clients' use of first-person pronouns has significantly reduced, showing that linguistic changes could be employed to monitor and understand the effects of therapy.

Apart from evaluating treatment progress and outcome, researchers have also attempted to use linguistic patterns to predict treatment outcomes. Huston and colleagues (2019) found that fewer past-oriented and negation words and more positive emotion words at an early point in treatment predicted better treatment outcomes. In another study that adopted a web-based depression treatment in the form of an online course targeted at depressed youth, researchers noted that increased use of "discrepancy" words

such as “should” during treatment predicted better learning of the online course (Van der Zanden et al., 2014). The use of linguistic markers in predicting treatment outcomes of depression is still limited; however, given the results of the previous studies, it seems promising to continue exploring the association between linguistic markers and outcomes of various forms of depression treatment.

Importance of attending to context. While research has demonstrated how several linguistic features can be used to predict depression, their diagnostic value is likely to be highly context-dependent. For example, the pattern in pronoun use seems to be sensitive to the nature of the prompt that the speech is produced in response to. Jarrold and colleagues (2011) found that increased self-focus in speech was only observed when the question was self-focused, broad, and evaluative. According to the study (Jarrold et al., 2011, p. 693), one example of such questions would be, “in your work or career, have you accomplished most of the things that you wanted to accomplish? (If no) Why not? What’s gotten in the way? Are you doing anything about this?” The question is part of a structured interview assessing personality characteristics (Carmelli et al., 1991; Rosenman et al., 1964). This sensitivity toward a topic or genre is not limited to the self-focus aspect of depressed language. In a study involving individuals with MDD, in addition to finding an interaction between first-person pronoun use and memory type, that the use of “I” decreased when the memory being recalled was positive and increased when the memory being recalled was negative, Himmelstein and colleagues (2018) also found the same pattern for present focus and overall word count. Furthermore, Havigerová and colleagues found that signs of depression in language appear more frequently in language tasks stressing informal rather than formal language (e.g., writing a letter about the holidays) (2019), again suggesting the need for paying attention to the context in which the language sample was produced.

Indeed, humans make sophisticated use of context in almost every aspect of our life, with or without awareness. For example, our color perception is greatly influenced by the context, and for the same reason, our depth perception is incredibly malleable, too. Returning to language, we rely heavily on context to interpret the meaning and make predictions. Consider the following two sentences:

(1) a. He is *running* to the building.

b. He is *running* for office.

The meaning of the word *running* differs in the two sentences in (1), clearly due to the words surrounding them. The influence of context is by no means restricted to the sentence level. Consider the following two exchanges:

(2) a. Do you want to order more food?

b. *I'm good, thank you.*

(3) a. How are you?

b. *I'm good, thank you.*

The sentences in (2) and (3) show how the meaning of the second sentences in the sequence, (2b) and (3b), changes as a result of context, demonstrating that the influence of context can certainly spill over sentence boundaries. Next, importantly, adjacent context is not the only thing we draw inferences from. Consider the above example, now with a couple of additional lines as shown in sentences (4) and (5):

(4) a. Do you want to order more food?

b. *I'm good.*

c. *How about you?*

d. *I'm good, thank you.*

(5) a. How are you?

b. *I'm good.*

c. *How about you?*

d. *I'm good, thank you.*

In this example, we would still interpret the sentences in (4d) and (5d) differently even though the two sentences preceding the response are the same across two exchanges. Above is just a crude demonstration of how much context matters when we use language. In reality, the influence of context is much more complicated and subtle. As such, language use is not only highly idiosyncratic but also incredibly context-

dependent, and one has to consider the context when trying to understand the individual differences in language use (especially when the size of the language sample is limited).

Limitations in previous methodology and technological advancements. When language first became a subject of interest, the examination of linguistic patterns relied largely on manual rating, which is labor-intensive and time-consuming. Furthermore, the inter-rater reliability could be a concern at times. Computerized approaches promised improvements over manual ratings. Early applications relied on pre-determined dimensions started. In the existing literature, one of the most widely used tools for computerized text analysis is the Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007). Most of the studies mentioned above in the section describing depressed language used LIWC (e.g., Himmelstein et al., 2018; Huston et al., 2019). LIWC is a powerful tool, but at the same time, it has obvious limitations – the analyses that LIWC enables are mostly restricted to the lexical level, which prevents it from detecting more abstract linguistic patterns given by phrases and sentences. As a result, previous research that studied depressed language has looked chiefly at individual features that were isolated from the context rather than holistically assessing a linguistic production in its entirety.

More recently, advances in big data have enabled the approach to move beyond the simple lexical items and examine patterns in linguistic behaviors more precisely and at a much greater scale than what could be achieved before (Johns et al., 2020). One important aspect of recent advancements is that computer-based models now have unprecedented ability to use context while interpreting, comparing, and predicting human language. The invention of the transformer (Vaswani et al., 2017), the neural network architecture underlying most of the best-performing large language models (LLMs), permitted this technological advancement. The LLMs used in the current project all use the transformer architecture.

Transformer architectures have three key characteristics: 1) positional encoding, 2) attention, and 3) self-attention. Positional encoding allows the word order information to be stored in the data itself rather than just a sequence in which the model receives the individual elements of the data, thus giving

the model the ability to learn the importance of word order from data. In addition, the ability to feed the model with a large amount of text with positional information encoded within, as opposed to having the model process text sequentially, greatly improves training efficiency and allows the model to be trained with larger amounts of data. Attention and self-attention are different but related notions – while attention allows the model to use information from any part of the *input* sequence during the *output* process, self-attention allows the model to access information from the *input* sequence during the *input* process, thus resulting in a more nuanced internal representation of language. Together, these three features of the transformer architecture give it the unprecedented ability to handle large amounts of training data and to use context holistically rather than relying on individual features.

Using the transformer architecture, language models like Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020) are able to outperform previous models on various text-based tasks. Notably, T5's performance correlates highly with human ratings on tasks like similarity judgment. In addition to the architectural improvement, researchers have also proposed innovative training and finetuning strategies that allow the resulting model to have strong ability to generalize – the ability to perform well in tasks that the model had limited experience with or had not encountered before. Finetuned LAnguage Net (FLAN; Wei et al., 2022) is one such method that can significantly improve models' performance on unseen tasks.

Given the recent advancements in language technology, we now have access to models that can perform various text-based tasks, seen or unseen, allowing for methodological shifts in research involving language data.

The Present Study

The present study examines if speech samples collected at an early stage of treatment have the potential to uncover subtypes within depression and whether these subtypes can inform us about treatment

outcomes. Specifically, we capitalize on the recent advancements in language technology and propose an LLM-based approach that disentangles the effect of context from idiosyncratic patterns of language use. The current project focused on identifying clusters in a group of clinically depressed individuals based on their speech in Cognitive Behavioral Therapy (CBT) sessions**.** Towards this goal, we first confirmed the LLM's sensitivity to context while predicting linguistic productions in connected speech. Next, we explored if the predictability of one's linguistic production under different contexts provides adequate signal for profiling one's language use pattern. We then examined if patterns of similarity imply reliable clusters and, if so, identify the linguistic features associated with those clusters. Lastly, we explored whether the cluster results were predictive of remission status.

Our key predictions were that the LLM we used would be sensitive to context, allowing us to observe different similarities between the predicted and the original production when the model was provided with different contexts – narratives of different people. We predicted that these similarities would allow us to identify different clusters of participants. We further expected that the clustering results would be predictive of the probability of remission.

Method

Sample

The current study focuses on a sample of clinically depressed individuals who were previously enrolled in a study that aimed at identifying predictors of treatment outcomes in major depressive disorder (MDD) (Dunlop et al., 2017). The 2017 study recruited 344 adults aged 18 to 65 who met DSM-IV criteria for MDD and were treatment-naïve, of which 115 were randomly assigned to receive Cognitive Behavioral Therapy (CBT) during the first 12 weeks (16 50-minute sessions: 2 sessions per week for the first 4 weeks, followed by 8 weekly sessions), 114 received escitalopram (10 – 20 mg per day) and 115 received duloxetine (30 – 60 mg per day). The present study sample was from the CBT group. Speech samples were not collected for the pharmacotherapy groups. Of the 115 participants in the CBT group, approximately one-third were drawn from a site where Spanish was the first language of the majority of

the participants. The speech samples from this group were not used in the current study. English language samples included 50 1st sessions of CBT. Remission status (remission, non-remission or early termination) was known for each of these participants and used as an outcome variable in the current study.

Data Processing

To prepare the session recordings for subsequent analyses, we developed a transcription pipeline using Azure, a HIPAA-compliant cloud computing service available through Microsoft. Speaker partitioning was also enabled through Azure's diarization libraries. To ensure data confidentiality, session recordings were preprocessed before being submitted to the transcription pipeline. Preprocessing involved the following steps:

1. All session recordings were converted from video to audio format.
2. Recordings were split into segments using the *split_on_silence* method in *Pydub*, a Python package containing methods for manipulating audios. The *split_on_silence* method identifies the silent sections in an audio to minimize the chance of splitting occurring mid-word. The resulting segments (2582 segments in total) are about 1 minute long each.
3. All segments were named with a random letter-number sequence ("2V8V4V6J.mp3") and the original order of the segments was stored locally.
4. The renamed segments were uploaded to a private Azure Blob, a system for secure data storage at scale for cloud-based computing. To use the batch transcription feature of Azure, this step was necessary.

All 2582 segmented recordings were run through the transcription pipeline. The pipeline outputted transcriptions as *.json* files. We extracted the transcribed speech and speaker information from each *.json* file. Importantly, the transcription pipeline outputted speaker information as "Speaker 1" and "Speaker 2" (if a second speaker is present), and we needed to determine which speaker was the therapist and which was the client in each segment before concatenating the transcripts back together. To determine whether

the therapist or the client was “Speaker 1” in each segment, a Python script was written to facilitate the identification process. The script ran through all transcripts, displaying the content of each transcript while playing the corresponding audio segment, and as soon as a human judge identified “Speaker 1” (pressing “T” for therapist and “C” for client), the program automatically updated the speaker column of the transcript (from “Speaker 1”/ “Speaker 2” to “therapist”/ “client’) and moved on to the next transcript. After determining the speaker identity, the 2582 transcripts were merged back into 50 files, each containing transcribed speech and speaker information from a CBT session.

Data Analysis Plan

1) Confirming LLM’s sensitivity to context during prediction.

To confirm LLM’s sensitivity to context when predicting connected speech, we examined if similarity between the predicted and the original production changed as we fed the model different contexts. In the current study, we used FLAN-UL2 for language prediction and T5-11b for similarity calculation. Each time FLAN-UL2 predicted a response in a section of connected speech, we looked at the top 20 solutions generated by the model and got a similarity score between each solution and the original production. We then recorded the maximum similarity as the score assessing how well the model predicts a particular response. For each individual, FLAN-UL2 predicted the responses to each question included in the speech sample, and the average of how well the model predicted each response was calculated as a metric representing the overall predictability of the individual’s speech.

Having determined a measure for assessing the predictability of one’s speech, we then examined if changing the context used by the model during prediction resulted in changes in the model’s performance. Specifically, we compared the similarity scores obtained when FLAN-UL2 was provided with minimal context (two sentences preceding the question that the model is required to generate a response to) and the similarity scores obtained when FLAN-UL2 was provided with an additional 1000-word connect speech sample containing the speaker-of-interest’s speech. The context length (i.e., 1000

words) was selected based on two reasons: 1) it represented the amount of language data that the model could reliably process in a single chunk, and 2) it was long enough to capture one of the key parts of the CBT session, for example, when participants were given an opportunity to describe the potential causes of their depression.

2) Characterizing idiosyncratic patterns of language use.

Once we confirm that FLAN-UL2 sentence predictions changed with different contexts, we constructed a similarity matrix reflecting FLAN-UL2's ability to predict sentences given different contexts. The contexts consisted of the 1000-word speech samples from each client.

To allow for inter-speaker comparison in subsequent analyses, we took into consideration FLAN-UL2's baseline predictability, which was based on the similarity between the sentences generated by the model and those that were actually produced by the patients when the prediction was not preceded by any narrative. We updated each similarity score in the matrix by subtracting the baseline similarity score from it.

3) Clustering and between-cluster feature analysis.

To identify possible subgroups in the dataset, we performed hierarchical cluster analysis. Hierarchical clustering does not require pre-specification of the number of clusters, allowing the technique to be used to determine not only category membership but also the number of clusters. To assess whether the clustering solution was stable, we examined the clustering solutions obtained when different dissimilarity metrics were used. If the resulting clustering solutions are similar, it suggests the clustering solutions are relatively stable. The validity of the solutions was also measured using the cophenetic correlation, which assesses how well a clustering solution preserves the pairwise distances in the original unmodeled data. Finally, we determine the optimum number of clusters using the elbow method and the silhouette analysis**.**

To better understand the clusters obtained, characteristic linguistic features were identified for each cluster. Semantic and syntactic linguistic features were extracted using the language analysis toolkit *spaCy*, a Python library for natural language processing. Chi-square tests were performed to determine if the observed frequency of linguistic features (semantic or syntactic) across the clusters differed from that expected based on the marginal sums. Subsequently, we looked at the adjusted residuals to identify the features that were characteristic of each cluster.

4) Exploring if the clustering results predict remission status.

Chi-square tests were conducted to determine whether the different clusters were related to remission status. Specifically, we examined if the remission status composition in each cluster differed from the base rate observed in the dataset. Such an analysis allowed us to evaluate the potential clinical utility of the clusters.

Results

The results were as predicted. Firstly, LLMs predicted sentences better when preceded by a narrative than when not preceded by a narrative, establishing that LLMs use context to generate predictions. The impact of these preceding narratives differed for different participants, consistent with the hypothesis that context effects can be used to measure the similarity of two speakers' speech. The resulting patterns of similarity were subjected to hierarchical clustering, which revealed an underlying natural partitioning of the participants. Participants were found to fall into three groups. Also, as expected, the members in each group shared certain syntactic and semantic trends. Intriguingly, the clusters were indicative of the probability of remission.

Confirming LLM's Sensitivity to Context During Prediction

We examined if the similarity scores obtained under the Minimal Context condition (*Mean* = 1.75, *SD* = 0.58) differed from those obtained under the 1000-word Context condition (*Mean* = 4.52, *SD*

= 0.43). *Figure 1* shows a scatter plot of the similarity scores obtained with a 1000-word context and a 2-word context.

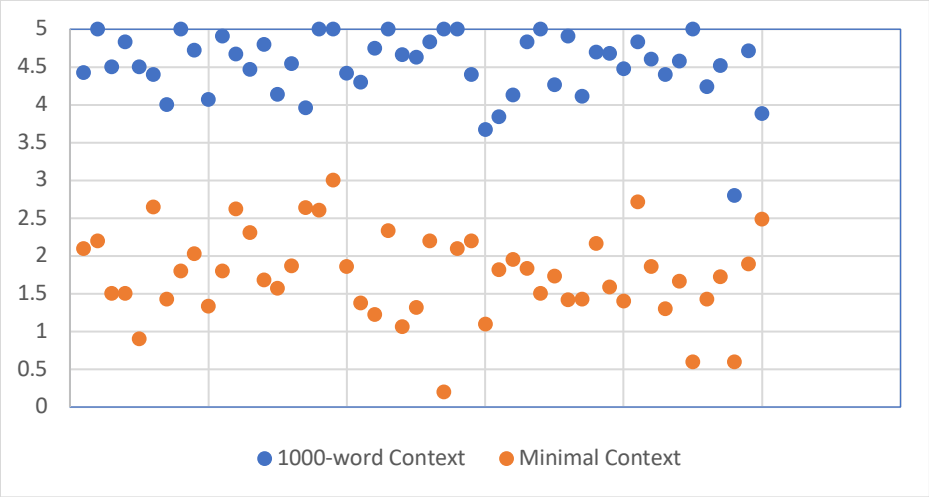

Figure 1. T5-11b similarity score between predicted response and original production in two conditions (1000-word Context and Minimal Context). Y-axis represents T5-11b similarity score.

A pared-samples t-test was conducted to compare the means. When predicting the same individual's speech, FLAN-UL2 produced predictions that were more similar to the original production in the 1000-word Context condition than in the Minimal Context condition, $t(49) = 30.31$, $p < 0.05$.

Natural Partitioning of Participants

We conducted agglomerative hierarchical clustering using a constant version of the Lance-Williams formula to conceptualize the locations of clusters and Euclidean Distance or Manhattan Distance to measure the dissimilarity between data points and clusters. Though the internal ordering of data points and low-level clusters differed from one dendrogram to the other, the higher-level grouping remained unchanged, suggesting a stable clustering structure. The dendrograms showing the clustering structures when the two different dissimilarity measures were used can be found in the supplemental materials (*Figures S1.a* and *S1.b*).

The cophenetic correlations of the two clustering solutions are 0.78 (when the dissimilarity measure is Manhattan Distance) and 0.71 (when the dissimilarity measure is Euclidean Distance). Importantly, we used the Spearman rank correlation coefficient (as opposed to the usual Pearson correlation coefficient) when obtaining the cophenetic correlations as it is proposed to be less sensitive to the presence of outliers (Timofeeva, 2019). One usually considers the dendrogram to accurately portray the original data structure when the cophenetic correlation is greater than 0.8 (Whitehead, 2009). Since the solution obtained using the Manhattan Distance yielded greater cophenetic correlation, we continued analyses based on this solution. Given that the cophenetic correlation was close to 0.8 but did not reach 0.8, we would interpret our results with caution.

We used the elbow method and the silhouette analysis to determine the optimum number of clusters. The Elbow Plot and the Silhouette Plots can be found in the supplemental materials (*Figures S2* and *S3s*). The elbow method looks at the relationship between the total within-cluster sum of squares (WCSS) and the number of clusters. On an Elbow Plot, one usually finds the point after which the line starts looking flat – the increase of number of clusters no longer results in significant reduction in WCSS. The elbow method suggests a three-cluster solution. We then corroborate the elbow method with the silhouette analysis. A higher overall average silhouette score is usually indicative of a better partitioning, however, the Silhouette Plot is important for assessing the quality of the solution, too. When looking at a Silhouette Plot, one usually pays attention to whether each cluster's average silhouette score is greater than the overall average silhouette score, and whether all clusters' plot has mostly uniform thickness. Returning to our results, though the 2-cluster partitioning resulted in the highest overall average silhouette score, the two clusters have very uneven thicknesses in the plot (*Figure S3.a*), and the average silhouette width of the bottom cluster is far from reaching the average silhouette width. It was hard to choose between the 3-cluster partitioning and the 4-cluster partitioning based on the two criteria described above, thus, we consider the overall average silhouette width, which led to the selection of the 3-cluster solution. As such, both the elbow method and the silhouette analysis pointed to the 3-cluster solution. *Figure 2*

shows the dendrogram obtained using the Manhattan Distance with 3-cluster partitioning.

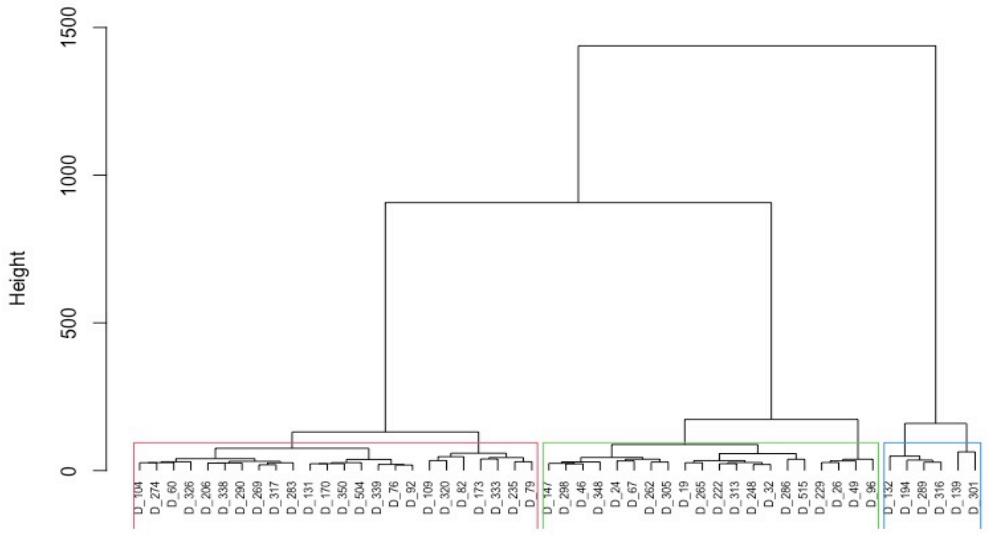

Figure 2. Clustering solution showing 3-cluster partitioning, using Manhattan Distance as dissimilarity measure.

Linguistic Characteristics of the Clusters

To better understand the clusters identified, we performed a linguistic feature analysis. Based on the clustering results, we grouped the clients into three groups (Cluster 1 – $N = 24$, Cluster 2 – $N = 20$, Cluster 3 – $N = 6$) and compared the distributions of linguistic features between the three clusters. The distributions of syntactic features (e.g., the use of past tense, the use of plural) differed significantly across the three clusters, X^2 (56, $N = 50$) = 284.09, $p < 0.05$. The distributions of semantic features (i.e., lemmatized words used by the speakers) also differed significantly across the three clusters, X^2 (674, $N = 50$) = 1183.71, $p < 0.05$. To better understand the nature of the three clusters, we examined the transcripts and conducted a contingency analysis which identified a list of key words that differed between the three clusters in terms of frequency. *Table 1* shows the linguistic features that were characteristic of each cluster.

Table 1

Linguistic Features More and Less Frequently Used by Individuals in Each Cluster

Cluster	1 (N = 24)		2 (N = 20)		3 (N = 6)	
Feature type	Syntactic	Semantic	Syntactic	Semantic	Syntactic	Semantic
More frequently used	**Past tense**	Son, bed, sleep, miss, year, night, class, **anxiety**, like, year, semester, mother, doctor, only, open	**Present tense**	Know, energy, pick, week, day, work, depressed, attack, watch, feeling, **enjoy**, issue, bill, continue, quit	**Past tense**	Sad, steal, supervisor, sick, know, enough, compare, relationship, people, part, timing, boyfriend, raise, anything, say
Less frequently used	Present tense	Tired, social, vacation, let, work, say, energy, know	Past tense	Class, **anxiety**, stress, event, compare, lose, miss, year, store, suppose, place, tell, see, son, control	/	Like, have

Importantly, individuals in Cluster 2 used present tense more frequently than individuals in the other two clusters, and individuals in the other two clusters used past tense more frequently than individuals in Cluster 2. Individuals in Cluster 1 talked more about anxiety, while those in Cluster 2 talked less about anxiety and stress. Additionally, individuals in Cluster 2 seemed to be the only ones talking about pleasurable activities, using words like “enjoy.” Another important distinction is that individuals in Cluster 2 emphasized content at the thematic and conceptual level, whereas individuals in Clusters 1 and 3 tended to focus on specific events and people. Speech samples from Cluster 3 also seemed less coherent. In all, we observed a clear distinction between Cluster 2 and the other two clusters.

Exploring if the Clustering Results Predict Remission Status

In *Figure 3*, we show the dendrogram with the 3-cluster partitioning together with the remission status of each individual. The outcome measure indicated that participants varied in their remission status, with 20 participants remitting, 24 participants not remitting, and 6 participants terminating early.

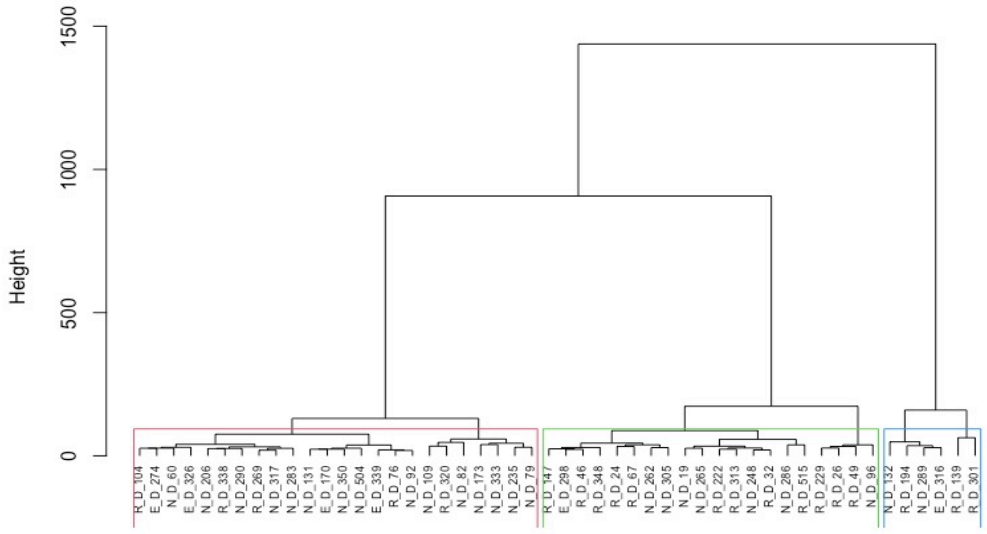

Figure 3. Clustering solution with remission status. The prefix "R_" indicates that the client reached full remission, "N_" indicates that the client had not reached full remission, and "E_" indicates that the client had terminated early before the planned sessions were completed.

A close inspection of the patient labels in *Figure 3* shows that remission rates differed across the clusters. In the leftmost cluster in *Figure 3*, Cluster 1 ($N = 24$), there are 5 remitters, 15 non-remitters, and 4 people who terminated early. The remission status composition differed significantly from the base rate observed in the sample, X^2 (2) = 15.29, $p < 0.05$. For the central cluster in *Figure 3*, Cluster 2, there are 12 remitters, 7 non-remitters, and 1 early terminator. These frequencies differed from the base rate observed in the sample, X^2 (2) = 19.41, $p < 0.05$. Finally, the rightmost cluster in *Figure 3*, Cluster 3, there are 3

remitters, 2 non-remitters, and 1 early terminator. The frequencies also differed from the base rates, X^2 (2) = 38.78, $p < 0.05$.

Critically, there were more remitters in Clusters 2 than in Cluster 1 and 3, there were more non-remitters in Cluster 1 than in Cluster 2 and 3. In all, these results suggest that Cluster 1 could be associated with non-remission and Cluster 2 could be associated with remission.

Discussion

In the current study, we confirmed LLM's sensitivity to context during prediction and capitalized on this feature of LLM to characterize individuals' speech production patterns. Predictability of one's speech as measured by the similarity between the predicted and the original production, served as signals for clustering analysis. Though the clustering solution was sub-optimal based on the cophenetic correlation, the clustering results were stable when different measures of dissimilarity were used. The elbow method and the silhouette analysis provided converging evidence for a three-cluster partitioning of the data. As such, we delved deeper into what was characteristic of the speech samples in each cluster while staying cognizant of the limits to our interpretation.

We observed a clear distinction between Cluster 2 and the other two clusters, noting several trends in linguistic features that differed across clusters. However, it is important to note that these trends were based on speech samples used in the clustering analysis – excerpts taken from the beginning of the sessions. As such, the linguistic characteristics described just now do not reflect language use in the entire session, and certainly do not reflect one's language use in general. Instead, they might be reflecting what was salient on one's mind at the start of the treatment.

Having attempted to characterize the linguistic-based heterogeneity within depression, we are interested in exploring whether the linguistic clusters are useful for predicting treatment outcomes in terms of remission status. The remission status composition of all three clusters differed from the base

rate observed in the overall sample, and a contingency analysis revealed that Cluster 1 contained more non-remitters and Cluster 2 contained more remitters. Given the cluster-specific linguistic characteristics discussed above, it makes sense that individuals in Cluster 2 were more likely to have remitted, considering that they were present-focused and were able to attend to pleasurable activities and identify patterns in their daily experience. Given that these speech samples were from the beginning of the treatment, it could be that this ability to attend to what is enjoyable and to extract important themes from daily experiences was a personal trait rather than a form of treatment gain.

Having emphasized the importance of context when establishing the method for this study, it is also important that we consider the context in which the speech samples are obtained when interpreting the findings. Specifically, the speech samples used are transcripts of CBT sessions, and we only included the first sessions in the current study. The choice of only including the first sessions as we start this line of research aligns with the goal of being able to predict treatment outcome early and optimize treatment plan early, but at the same time, only including the first sessions restricts the generalizability of our findings, in that speech samples obtained from a later point in treatment may not exhibit the same features. As such, it would be interesting to look at speech samples from later sessions and compare the clustering and feature analysis results to the current study's findings. It could be that cluster membership does not change much, but the linguistic features characterizing the clusters change. Alternatively, cluster membership may change in that individuals could be grouped into different clusters while the features characterizing each cluster remain the same as what we found using the first sessions. Figuring out which features are associated with which outcome at what point of the treatment is important for assessing the clinical utility of language as a biomarker.

The characteristics of the participants also have consequences for evaluating the external validity of a study. The study where we obtained our current sample (Dunlop et al., 2017) was a randomized control study. As such, in order to study the variables of interest in a controlled setting, the researchers had to include a number of exclusionary criteria while recruiting their participants. These criteria include

certain comorbidities, prior treatment history, lifetime use of certain medications, *et cetera*. The presence of stringent exclusionary criteria could result in the sample being more homogeneous than a typical sample of depressed population. As such, we should validate the current method with different samples and be cautious when generalizing the current findings. That said, given that language use exhibits sufficient sensitivity to be used as a subtyping tool in the current sample (a relatively more homogeneous sample), there is reason to believe that language use would have adequate discriminative power when we look at a more heterogenous sample.

It is also important to recognize that remission status is not the only treatment outcome measure. Symptom severity, symptom trajectories, functional recovery, and quality of life are among the other commonly examined outcome measures. To better understand the predictive value of language, we should also examine the relationship between linguistic-based clusters and the other available treatment outcome measures. Also, in addition to having a cut-off score or a fixed threshold, change from baseline is often used as well when describing treatment outcome. Relatedly, change in one's linguistic productions could be investigated in the future.

Finally, investigating the predictive value of language use on its own is likely not sufficient. Previous research suggests that a single predictor rarely predicts treatment outcome in depression successfully (Rost et al., 2023). As we develop reliable and valid measures that characterize one's language use, we could start combining linguistic-based measures with other available predictors of treatment outcome in depression and improve the current predictive models.

Conclusion

In the current study, we proposed an LLM-based approach that disentangles the effect of context from idiosyncratic patterns of language use, and examined if speech sample collected at early stage of treatment could be used to characterize the heterogeneity within depression. We discovered a three-cluster

partitioning in the sample and two of the three clusters seemed to indicate remission status. Although the generalizability of the current finding is limited given the specificity of the participants group and the speech sample, our finding suggests that language has the potential to be a subtyping tool and a predictor of treatment outcome in depression. Furthermore, the method we proposed in the study will allow us to efficiently apply future language models and speech samples as they become available.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Baddeley, J. L., Daniel, G. R., & Pennebaker, J. W. (2011). How Henry Hellyer's use of language foretold his suicide. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, *32*(5), 288–292. https://doi.org/10.1027/0227-5910/a000092

Baller, E. B., Kaczkurkin, A. N., Sotiras, A., Adebimpe, A., Bassett, D. S., Calkins, M. E., Chand, G. B., Cui, Z., Gur, R. E., Gur, R. C., Linn, K. A., Moore, T. M., Roalf, D. R., Varol, E., Wolf, D. H., Xia, C. H., Davatzikos, C., & Satterthwaite, T. D. (2021). Neurocognitive and functional heterogeneity in depressed youth. *Neuropsychopharmacology*, *46*(4), 783–790. https://doi.org/10.1038/s41386-020-00871-w

Beijers, L., Wardenaar, K. J., van Loo, H. M., & Schoevers, R. A. (2019). Data-driven biological subtypes of depression: Systematic review of biological approaches to depression subtyping. *Molecular Psychiatry*, *24*(6), 888–900. https://doi.org/10.1038/s41380-019-0385-5

Bernard, J. D., Baddeley, J. L., Rodriguez, B. F., & Burke, P. A. (2016). Depression, Language, and Affect: An Examination of the Influence of Baseline Depression and Affect Induction on Language. *Journal of Language and Social Psychology*, *35*(3), 317–326. https://doi.org/10.1177/0261927X15589186

Bueno-Notivol, J., Gracia-García, P., Olaya, B., Lasheras, I., López-Antón, R., & Santabárbara, J. (2021). Prevalence of depression during the COVID-19 outbreak: A meta-analysis of community-based studies. *International Journal of Clinical and Health Psychology*, *21*(1), 100196. https://doi.org/10.1016/j.ijchp.2020.07.007

Carmelli, D., McElroy, M. R., & Rosenman, R. H. (1991). Longitudinal changes in fat distribution in the Western Collaborative Group Study: A 23-year follow-up. *International Journal of Obesity*, *15*(1), 67–74.

Cuijpers, P., Stringaris, A., & Wolpert, M. (2020). Treatment outcomes for depression: Challenges and

opportunities. *The Lancet Psychiatry*, *7*(11), 925–927. https://doi.org/10.1016/S2215-0366(20)30036-5

Demiray, Ç. K., & Gençöz, T. (2018). Linguistic reflections on psychotherapy: Change in usage of the first person pronoun in information structure positions. *Journal of Psycholinguistic Research*, *47*(4), 959–973. https://doi.org/10.1007/s10936-018-9569-4

Dunlop, B. W., Kelley, M. E., Aponte-Rivera, V., Mletzko-Crowe, T., Kinkead, B., Ritchie, J. C., Nemeroff, C. B., Craighead, W. E., Mayberg, H. S., & for the PReDICT Team. (2017). Effects of Patient Preferences on Outcomes in the Predictors of Remission in Depression to Individual and Combined Treatments (PReDICT) Study. *American Journal of Psychiatry*, *174*(6), 546–556. https://doi.org/10.1176/appi.ajp.2016.16050517

Edwards, T., & Holtzman, N. S. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, *68*, 63–68. https://doi.org/10.1016/j.jrp.2017.02.005

Friedrich, M. J. (2017). Depression Is the Leading Cause of Disability Around the World. *JAMA*, *317*(15), 1517. https://doi.org/10.1001/jama.2017.3826

Gao, S., Calhoun, V. D., & Sui, J. (2018). Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neuroscience & Therapeutics*, *24*(11), 1037–1052. https://doi.org/10.1111/cns.13048

Greenberg, J., & Pyszczynski, T. (1986). Persistent high self-focus after failure and low self-focus after success: The depressive self-focusing style. *Journal of Personality and Social Psychology*, *50*(5), 1039–1044. https://doi.org/10.1037/0022-3514.50.5.1039

Groves, S. J., Douglas, K. M., & Porter, R. J. (2018). A Systematic Review of Cognitive Predictors of Treatment Outcome in Major Depression. *Frontiers in Psychiatry*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyt.2018.00382

Habert, J., Katzman, M. A., Oluboka, O. J., McIntyre, R. S., McIntosh, D., MacQueen, G. M., Khullar,

A., Milev, R. V., Kjernisted, K. D., & Chokka, P. R. (2016). Functional Recovery in Major Depressive Disorder: Focus on Early Optimized Treatment. *The Primary Care Companion for CNS Disorders*, *18*(5), 24746. https://doi.org/10.4088/PCC.15r01926

Harald, B., & Gordon, P. (2012). Meta-review of depressive subtyping models. *Journal of Affective Disorders*, *139*(2), 126–140. https://doi.org/10.1016/j.jad.2011.07.015

Havigerová, J. M., Haviger, J., Kučera, D., & Hoffmannová, P. (2019). Text-Based Detection of the Risk of Depression. *Frontiers in Psychology*, *10*, 513. https://doi.org/10.3389/fpsyg.2019.00513

Himmelstein, P., Barb, S., Finlayson, M. A., & Young, K. D. (2018). Linguistic analysis of the autobiographical memories of individuals with major depressive disorder. *PLoS ONE*, *13*(11). https://doi.org/10.1371/journal.pone.0207814

Huston, J., Meier, S., Faith, M., & Reynolds, A. (2019). Exploratory study of automated linguistic analysis for progress monitoring and outcome assessment. *Counselling & Psychotherapy Research*, *19*(3), 321–328. https://doi.org/10.1002/capr.12219

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *The American Journal of Psychiatry*, *167*(7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

Jarrold, W., Javitz, H. S., Krasnow, R., Peintner, B., Yeh, E., Swan, G. E., & Mehl, M. (2011). Depression and self-focused language in structured interviews with older men. *Psychological Reports*, *109*(2), 686–700. https://doi.org/10.2466/02.09.21.28.PR0.109.5.686-700

Johns, B. T., Jamieson, R. K., & Jones, M. N. (2020). The continued importance of theory: Lessons from big data approaches to language and cognition. In *Big data in psychological research* (pp. 277–295). American Psychological Association. https://doi.org/10.1037/0000193-013

Kauschke, C., Mueller, N., Kircher, T., & Nagels, A. (2018). Do Patients With Depression Prefer Literal or Metaphorical Expressions for Internal States? Evidence From Sentence Completion and Elicited Production. *Frontiers in Psychology*, *9*, 1326. https://doi.org/10.3389/fpsyg.2018.01326

Kim, K., Choi, S., Lee, J., & Sea, J. (2019). Differences in linguistic and psychological characteristics between suicide notes and diaries. *Journal of General Psychology*, *146*(4), 391–416. https://doi.org/10.1080/00221309.2019.1590304

Kraus, C., Kadriu, B., Lanzenberger, R., Zarate Jr., C. A., & Kasper, S. (2019). Prognosis and improved outcomes in major depression: A review. *Translational Psychiatry*, *9*, 127. https://doi.org/10.1038/s41398-019-0460-3

Lamers, F., Burstein, M., He, J., Avenevoli, S., Angst, J., & Merikangas, K. R. (2012). Structure of major depressive disorder in adolescents and adults in the US general population. *The British Journal of Psychiatry*, *201*(2), 143–150. https://doi.org/10.1192/bjp.bp.111.098079

Laursen, T. M., Musliner, K. L., Benros, M. E., Vestergaard, M., & Munk-Olsen, T. (2016). Mortality and life expectancy in persons with severe unipolar depression. *Journal of Affective Disorders*, *193*, 203–207. https://doi.org/10.1016/j.jad.2015.12.067

Lyons, M., Aksayli, N. D., & Brewer, G. (2018). Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior*, *87*, 207–211. https://doi.org/10.1016/j.chb.2018.05.035

Mariani, R., Di Trani, M., Negri, A., & Tambelli, R. (2020). Linguistic analysis of autobiographical narratives in unipolar and bipolar mood disorders in light of multiple code theory. *Journal of Affective Disorders*, *273*, 24–31. https://doi.org/10.1016/j.jad.2020.03.170

Moreno-Agostino, D., Wu, Y.-T., Daskalopoulou, C., Hasan, M. T., Huisman, M., & Prina, M. (2021). Global trends in the prevalence and incidence of depression:a systematic review and meta-analysis. *Journal of Affective Disorders*, *281*, 235–243. https://doi.org/10.1016/j.jad.2020.12.035

Murray, C. J. L., & World Health Organization (Eds.). (2002). *Summary measures of population health: Concepts, ethics, measurement, and applications*. World Health Organization.

Newell, E. E., McCoy, S. K., Newman, M. L., Wellman, J. D., & Gardner, S. K. (2018). You Sound So Down: Capturing Depressed Affect Through Depressed Language. *Journal of Language and Social Psychology*, *37*(4), 451–474. https://doi.org/10.1177/0261927X17731123

Paris, J. (2014). The Mistreatment of Major Depressive Disorder. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, *59*(3), 148–151.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic inquiry and word count: LIWC. Austin, TX.

Pulverman, C. S., Lorenz, T. A., & Meston, C. M. (2015). Linguistic changes in expressive writing predict psychological outcomes in women with history of childhood sexual abuse and adult sexual dysfunction. *Psychological Trauma: Theory, Research, Practice, and Policy*, *7*(1), 50–57. https://doi.org/10.1037/a0036462

Quinn, C. R., Rennie, C. J., Harris, A. W. F., & Kemp, A. H. (2014). The impact of melancholia versus non-melancholia on resting-state, EEG alpha asymmetry: Electrophysiological evidence for depression heterogeneity. *Psychiatry Research*, *215*(3), 614–617. https://doi.org/10.1016/j.psychres.2013.12.049

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv:1910.10683 [Cs, Stat]*. http://arxiv.org/abs/1910.10683

Rehm, J., & Shield, K. D. (2019). Global Burden of Disease and the Impact of Mental and Addictive Disorders. *Current Psychiatry Reports*, *21*(2), 10. https://doi.org/10.1007/s11920-019-0997-0

Rodgers, S., Grosse Holtforth, M., Müller, M., Hengartner, M. P., Rössler, W., & Ajdacic-Gross, V. (2014). Symptom-based subtypes of depression and their psychosocial correlates: A person-centered approach focusing on the influence of sex. *Journal of Affective Disorders*, *156*, 92–103. https://doi.org/10.1016/j.jad.2013.11.021

Rosenman, R. H., Friedman, M., Straus, R., Wurm, M., Kositchek, R., Hahn, W., & Werthessen, N. T. (1964). A PREDICTIVE STUDY OF CORONARY HEART DISEASE. *JAMA*, *189*, 15–22. https://doi.org/10.1001/jama.1964.03070010021004

Rost, N., Binder, E. B., & Brückl, T. M. (2023). Predicting treatment outcome in depression: An

introduction into current concepts and challenges. *European Archives of Psychiatry and Clinical Neuroscience*, *273*(1), 113–127. https://doi.org/10.1007/s00406-022-01418-4

Rude, S. S., Gortner, E.-M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, *18*(8), 1121–1133. https://doi.org/10.1080/02699930441000030

Smirnova, D., Cumming, P., Sloeva, E., Kuvshinova, N., Romanov, D., & Nosachev, G. (2018). Language Patterns Discriminate Mild Depression From Normal Sadness and Euthymic State. *Frontiers in Psychiatry*, *9*, 105. https://doi.org/10.3389/fpsyt.2018.00105

Smirnova, D., Romanov, D., Sloeva, E., Kuvshinova, N., Cumming, P., & Nosachev, G. (2019). Language in Mild Depression: How It Is Spoken, What It Is About, and Why It Is Important to Listen. *Psychiatria Danubina*, *31*, 427–433.

Sonnenschein, A. R., Hofmann, S. G., Ziegelmayer, T., & Lutz, W. (2018). Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy. *Cognitive Behaviour Therapy*, *47*(4), 315–327. https://doi.org/10.1080/16506073.2017.1419505

ten Have, M., Lamers, F., Wardenaar, K., Beekman, A., de Jonge, P., van Dorsselaer, S., Tuithof, M., Kleinjan, M., & de Graaf, R. (2016). The identification of symptom-based subtypes of depression: A nationally representative cohort study. *Journal of Affective Disorders*, *190*, 395–406. https://doi.org/10.1016/j.jad.2015.10.040

Timofeeva, A. (2019). Evaluating the robustness of goodness-of-fit measures for hierarchical clustering. *Journal of Physics: Conference Series*, *1145*(1), 012049. https://doi.org/10.1088/1742-6596/1145/1/012049

Trifu, R. N., Nemeş, B., Bodea-Hategan, C., & Cozman, D. (2017). Linguistic indicators of language in major depressive disorder (MDD) An evidence based research. *Journal of Evidence-Based Psychotherapies*, *17*(1), 105–128. https://doi.org/10.24193/jebp.2017.1.7

Uher, R., Tansey, K. E., Malki, K., & Perlis, R. H. (2012). Biomarkers predicting treatment outcome in

depression: What is clinically significant? *Pharmacogenomics*, *13*(2), 233–240. https://doi.org/10.2217/pgs.11.161

Van der Zanden, R., Curie, K., Van Londen, M., Kramer, J., Steen, G., & Cuijpers, P. (2014). Web-based depression treatment: Associations of clients' word use with adherence and outcome. *Journal of Affective Disorders*, *160*, 10–13. https://doi.org/10.1016/j.jad.2014.01.005

van Loo, H. M., de Jonge, P., Romeijn, J.-W., Kessler, R. C., & Schoevers, R. A. (2012). Data-driven subtypes of major depressive disorder: A systematic review. *BMC Medicine*, *10*(1), 156. https://doi.org/10.1186/1741-7015-10-156

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Vicent-Gil, M., Portella, M. J., Serra-Blasco, M., Navarra-Ventura, G., Crivillés, S., Aguilar, E., Palao, D., & Cardoner, N. (2020). Dealing with heterogeneity of cognitive dysfunction in acute depression: A clustering approach. *Psychological Medicine*, 1–9. https://doi.org/10.1017/S0033291720001567

Wadsworth, M. E., Hudziak, J. J., Heath, A. C., & Achenbach, T. M. (2001). Latent Class Analysis of Child Behavior Checklist Anxiety/Depression in Children and Adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(1), 106–114. https://doi.org/10.1097/00004583-200101000-00023

Watkins, E., & Teasdale, J. D. (2004). Adaptive and maladaptive self-focus in depression. *Journal of Affective Disorders*, *82*(1), 1–8. https://doi.org/10.1016/j.jad.2003.10.006

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). *Finetuned Language Models Are Zero-Shot Learners* (arXiv:2109.01652). arXiv. http://arxiv.org/abs/2109.01652

Whitehead, H. (2009). SOCPROG Programs: Analysing Animal Social Structures. *Behavioral Ecology and Sociobiology*, *63*(5), 765–778.

Zimmerman, M., Ellison, W., Young, D., Chelminski, I., & Dalrymple, K. (2015). How many different

ways do patients meet the diagnostic criteria for major depressive disorder? *Comprehensive Psychiatry*, *56*, 29–34. https://doi.org/10.1016/j.comppsych.2014.09.007

Supplemental Materials

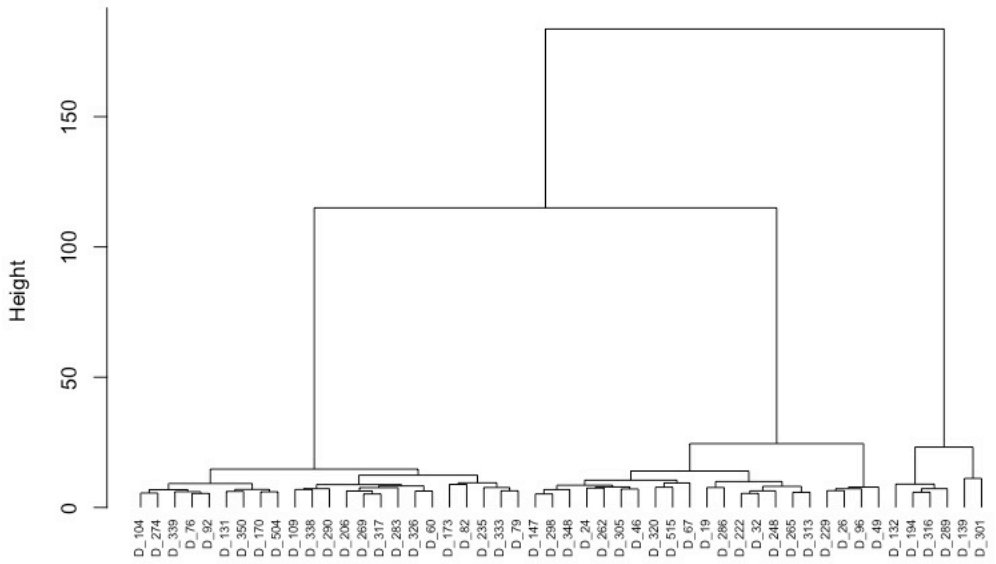

Figure S1.a. Clustering solution using Euclidean Distance as the dissimilarity measure.

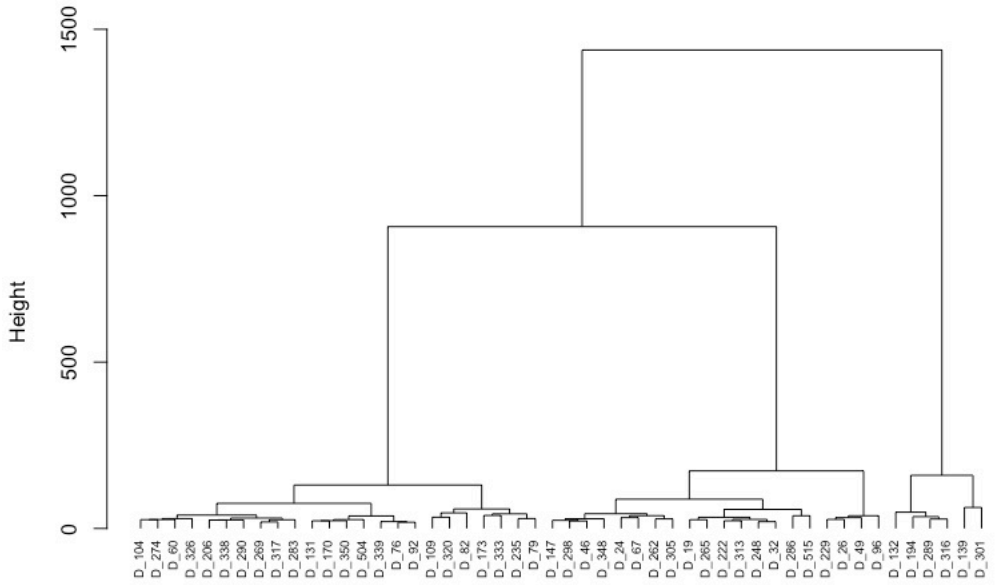

Figure S1.b. Clustering solution using Manhattan Distance as the dissimilarity measure.

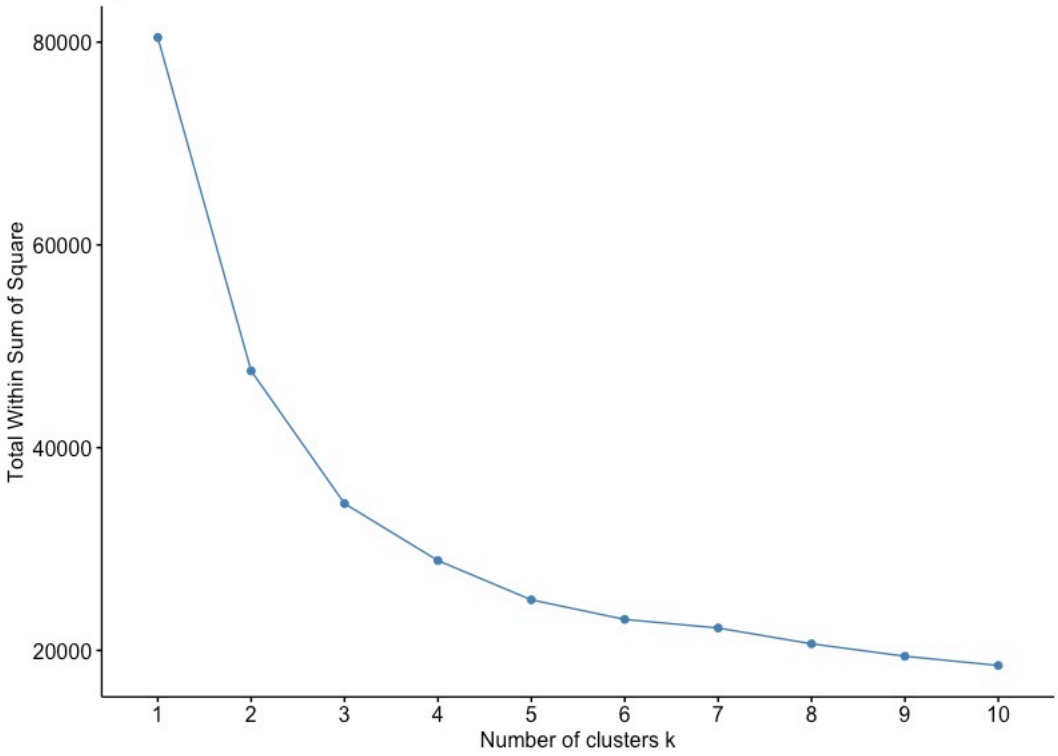

Figure S2. Elbow Plot showing total within-cluster sum of squares as number of clusters k changes.

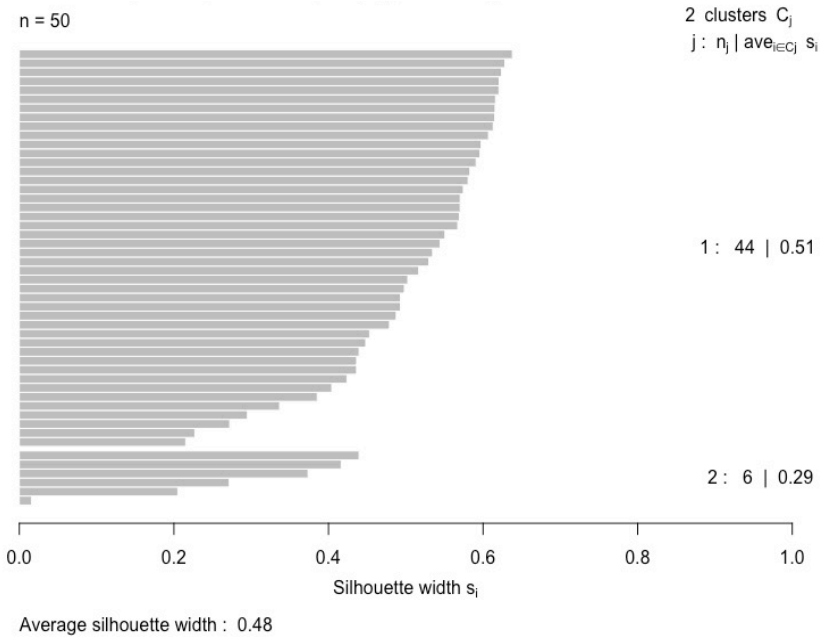

Figure S3.a. Silhouette Plot when number of clusters equals 2.

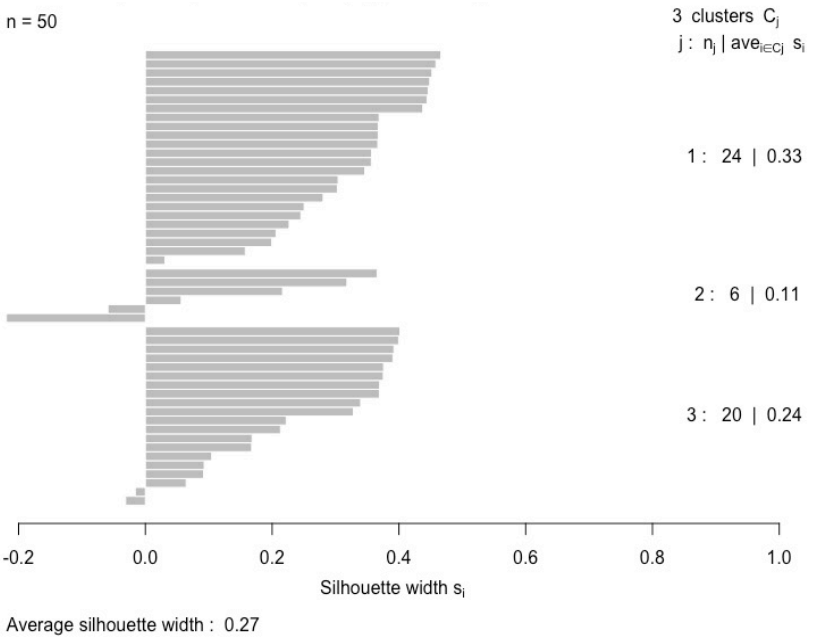

Figure S3.b. Silhouette Plot when number of clusters equals 3.

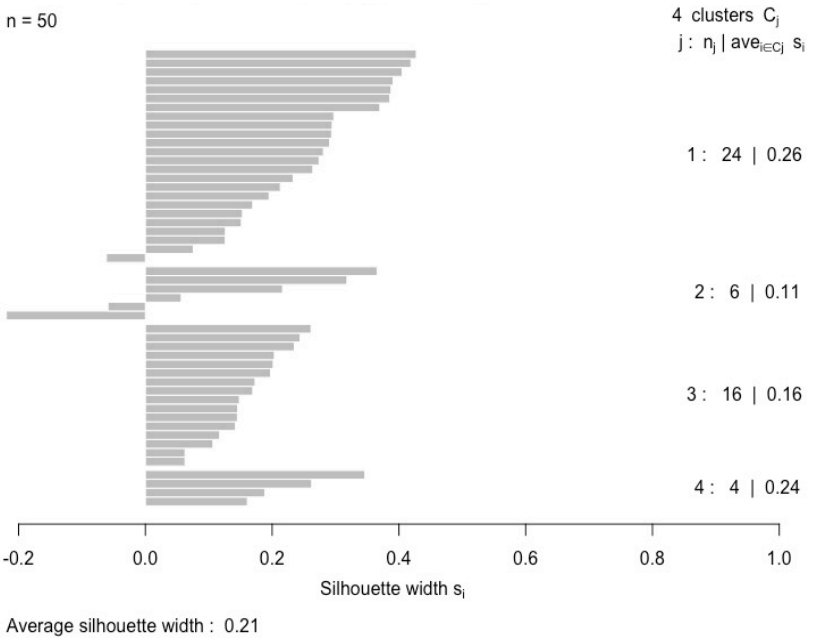

Figure S3.c. Silhouette Plot when number of clusters equals 4.