Linguistic Complexity Loss in Text-Based Therapy

Jason Wei♣[*] **Kelly Finn**🐙△ **Emma Templeton**🐙 **Thalia Wheatley**🐙 **Soroush Vosoughi**♣
Dartmouth College, Hanover, NH
♣Department of Computer Science △Neukom Institute
🐙Department of Psychological and Brain Sciences
{jason.20, kelly.r.finn, emma.m.templeton.gr,
thalia.p.wheatley, soroush}@dartmouth.edu

Abstract

The complexity loss paradox, which posits that individuals suffering from disease exhibit surprisingly predictable behavioral dynamics, has been observed in a variety of both human and animal physiological systems. The recent advent of online text-based therapy presents a new opportunity to analyze the complexity loss paradox in a novel operationalization: linguistic complexity loss in text-based therapy conversations.

In this paper, we analyze linguistic complexity correlates of mental health in the online therapy messages sent between therapists and 7,170 clients who provided 30,437 corresponding survey responses on their anxiety. We found that when clients reported more anxiety, they showed reduced lexical diversity as estimated by the moving average type-token ratio. Therapists, on the other hand, used language of higher reading difficulty, syntactic complexity, and age of acquisition when clients were more anxious. Finally, we found that clients, and to an even greater extent, therapists, exhibited consistent levels of many linguistic complexity measures. These results demonstrate how linguistic analysis of text-based communication can be leveraged as a marker for anxiety, an exciting prospect in a time of both increased online communication and increased mental health issues.

1 Introduction

The complexity loss paradox (Goldberger, 1997) posits that individuals suffering from a wide range of illnesses tend to exhibit surprisingly periodic and predictable dynamics in their behavior, even though the diseases themselves are often called *dis*-orders. The paradox exists in patterns of behavior from diving in penguins (Cottin et al., 2014) to social interactions in chimpanzees (Alados and Huffman, 2000). In humans, the paradox has been observed in physiological systems from the indistinguishable tremors of Parkinsonian patients (Parker et al., 2018) to the cyclic oscillations of white blood cell counts in leukemia patients (Malhotra and Salam, 1991), but how the paradox manifests in one of our most important behavioral outputs—language—has not been well-studied.

	Dataset	
	Exploratory	Confirmatory
Messages	2.6 million	0.7 million
Survey responses	24,287	6,150
Clients	5,736	1,434
Therapists	1,608	889
†Survey responses / client	4.23	4.29
†Client text (words) / survey	1259	1295
†Therapist text (words) / survey	796	804
Median survey score (0-21)	8	8
Median time between surveys	21 days	21 days

Table 1: Descriptive statistics for Talkspace online therapy conversations dataset. † indicates mean.

In what form could the complexity loss paradox manifest in language? A line of psycholinguistics research, starting from the 1970s, has shown that the words people use can reveal important aspects of their mental health (Pennebaker et al., 2003). For instance, vague and qualified speech can predict depression (Andreasen and Pfohl, 1976), diversity of word usage can indicate stress in interviews (Höweler, 1972), and other work has found that lexical choices correlate with aphasia (Wachal and Spreen, 1973) and suicide (Pestian et al., 2012).

In today's digital era, people suffering from mental illness have increasingly sought therapy services online, which can be more accessible than traditional clinicians' offices (Hull et al., 2018). Many online platforms serve a large number of clients through text-based therapy, and so these conversations (when anonymized and used with consent) are well-suited for computational analysis. Prior work has already used computational methods to predict symptom severity (Howes et al., 2014), measure

[*] Now AI Resident at Google.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4450–4459
June 6–11, 2021. ©2021 Association for Computational Linguistics

counseling quality (Pérez-Rosas et al., 2018, 2019), and used topic models to support counselors during conversations (Dinakar et al., 2015).

In this paper, we explore the complexity loss paradox in online therapy conversations of patients with anxiety. Whereas much recent work using NLP to find linguistic indicators of mental health has turned to social media data (Coppersmith et al., 2014; Benton et al., 2017), which is collected in a non-clinical context and may be unreliable, here we analyze a large-scale dataset of therapy conversations comprising 7,170 clients who sent more than three-million messages and answered 30,437 surveys about their mental health. Moreover, therapy is a dynamic activity between clients and therapists, and so compared with related work that focuses solely on linguistic patterns of counselors (Althoff et al., 2016; Zhang et al., 2019; Lee et al., 2019), we investigate linguistic complexity in both clients and therapists. What linguistic complexity patterns in the language of clients and therapists during therapy reflect client mental health?

2 Dataset

Talkspace. In this work, we study text-based messages from Talkspace, an online therapy platform with thousands of licensed therapists serving more than one-million users (Talkspace, 2020). Anyone seeking therapy, henceforth *clients*, can sign up for a Talkspace plan and get matched with a licensed *therapist* who will respond 5× a week through a chat room accessible by clients 24-7.

To assess client mental health, counselors send surveys to clients at periodic intervals (on average, every three weeks). Clients with different mental health conditions receive different surveys, with the most frequent surveys gauging anxiety and depression. In this work, we focus on anxiety, which clients self-reported using the Generalized Anxiety Disorder 7-item scale (Spitzer et al., 2006). Clients answer how often in the last two weeks they were bothered by certain problems (e.g., trouble relaxing or feeling afraid as if something awful might happen) on a scale from 0-3 (0: not at all sure, 3: nearly every day). Answers for the seven questions summed to a total score from 0-21, with 0 as the least anxious and 21 as the most anxious.

Dataset. Our dataset (summarized by Table 1) contains messages between clients and therapists on Talkspace sent between January 2016 and July 2019. We filtered these messages for those between therapists and adult clients for which clients had completed at least 6 weeks of treatment and responded to least 2 anxiety surveys that each had messages of at least 50 words within the week prior.

We take several precautions to reduce the probability of Type I errors. Upon receiving the dataset, we first followed Fafchamps and Labonne (2016) and split the dataset by client into an *exploratory* dataset (80%) and a *confirmatory* dataset (20%). We used the exploratory dataset for running analyses and making design decisions, and then preregistered our analyses and expected results before accessing the confirmatory dataset to perform a full replication of experiments. As such, throughout the paper, we report numbers from the exploratory dataset, but only indicate statistical significance that holds on both the exploratory and confirmatory datasets. To further reduce potential false positives, because we run $k{=}48$ tests for given data, we apply the Bonferroni correction (Cabin and Mitchell, 2000) and divide the traditional $\alpha{=}0.05$ by k so that we only consider statistical significance when $p < 0.001$.

Data Privacy. All patients and clinicians gave consent to the use of their data in a de-identified, aggregate format as part of the user agreement before they begin using the platform and can opt out at any time by informing their therapist or by contacting support. Study procedures were approved as exempt by the our institution's Institutional Review Board (IRB).

Transcripts were de-identified algorithmically via a HIPAA-compliant interface by anonymizing all proper nouns, places, persons, and other nominal features of language. All information related to forms of contact are also removed, including emails, phone numbers, addresses, though these were infrequently found in the interaction between therapists and patients.

3 Linguistic Complexity Measures

Linguistic complexity is a multi-faceted topic for which there is no single agreed-upon measure for indexing complexity; instead, a toolbox of measures should be used to assess various linguistic features (Goldberger et al., 2002). In this work, we consider twelve well-known linguistic complexity measures, compiled from the work of Tsvetkov et al. (2016), Mccarthy and Jarvis (2010), and popular readability formulas. We group these twelve complexity measures into four broad categories:

lexical diversity (🌏), syntactic simplicity (🌳), readability (📖), and prototypicality (👶). We list these complexity measures below, and direct the involved reader to the Appendix for details.

1. 🌏 Moving Average Type Token Ratio (MATTR): We use the moving average type-token ratio (MATTR) (Covington and McFall, 2010)—for a given sequence of tokens, we slide a window of size $W = 50$ over all tokens with a stride of $s = 1$, compute TTR (#types / #tokens) for each of the windows, and output the average.

2. 🌏 HD-D: HD-D (McCarthy and Jarvis, 2007) measures the mean contribution that each type makes to the TTR of all possible combinations of text samples of size 35-50, where higher HD-D indicates greater lexical diversity.

3. 🌏 Measure of Textual Lexical Diversity (MTLD): MTLD (McCarthy, 2005) measures the mean length of word strings that maintain a criterion level of lexical variation.

4. 🌳 Dependency parse tree depth.

5. 🌳 Sentence length: words per sentence.

6. 📖 Dale-Chall readability score (Dale and Chall, 1948, 1995): texts with higher DCRS are supposed to be more challenging to read.

7. 📖 Coleman-Liau index (Coleman and Liau, 1975): approximates the U.S. grade level thought necessary to comprehend the text.

8. 📖 Flesch-Kincaid grade level (Kincaid et al., 1975): higher scores indicate material that is more challenging to read.

9. 👶 Age of acquisition (AoA): extracted from a database of crowd-sourced ratings of over 30 thousand words (Kuperman et al., 2012).

10. 👶 Concreteness: averaged word-level concreteness ratings on the scale from 1–5 (1 is most abstract, and 5 is most concrete) for 40 thousand English lemmas (Brysbaert et al., 2014).

11. 👶 Syllable count: average syllables per word.

12. Talkativeness: number of alphanumeric tokens for either client or therapist in a conversation, which we define as all messages in the one week period before a survey.

Effect	β	t	p
Weeks in therapy	-0.26	-18.31	$< 2\times10^{-16}$
GENDER			
Female	0.02	0.26	0.80
Gender Queer	0.40	2.23	0.026
Gender Variant	0.51	1.61	0.11
Male	-0.05	-0.53	0.60
Other	0.37	1.87	0.061
Transgender Female	0.42	1.81	0.070
Transgender Male	0.36	1.38	0.17
EDUCATION			
Associate's Degree	-0.05	-0.48	0.63
Bachelor's Degree	-0.04	-1.50	0.13
Doctoral Degree	-0.06	-0.483	0.63
High School	0.13	3.38	0.00073
Less than High School	0.42	2.77	0.0057
Master's Degree	0.05	0.77	0.44
Professional Degree	0.40	2.68	0.0074
Some College No Degree	0.17	2.60	0.0094
AGE			
18–25	0.08	2.20	0.028
26–35	-0.04	-1.41	0.16
36–49	-0.11	-2.92	0.0035
50+	-0.36	-5.78	8×10^{-9}

Table 2: Demographic predictors of client anxiety that were controlled for in our linear mixed model analysis. Positive β indicates positive correlation with reported anxiety, and negative β indicates negative correlation with anxiety.

4 Complexity Correlates of Anxiety

We investigate how measures of linguistic complexity varied with reported client anxiety. For the 5,736 clients in the exploratory dataset, we retrieve all messages sent within one week prior to an anxiety survey response—henceforth *conversations*—totaling 24,287 conversation-survey pairs.

For all conversation-survey pairs, we compute a value $\mathcal{C}^m$ for each complexity measure m and both clients and therapist messages in that conversation. We then observe how each complexity measure changes with client anxiety (normalized for demographic variables) using a linear mixed model (Galecki and Burzykowski, 2013), which models random effects (variables that account for differences across individuals) as well as fixed effects in a general linear model. We predict anxiety using $\mathcal{C}^m$ as a fixed effect, and, to control for demographic variables and individual differences, we also model time in therapy, gender, education, and age as fixed effects, and include therapist ID and client ID as random effects, with time as a random slope on client ID. Table 2 shows these demographic variables and their effects that we control for. As we are interested in the effect of each complexity measure on anxiety, we run this

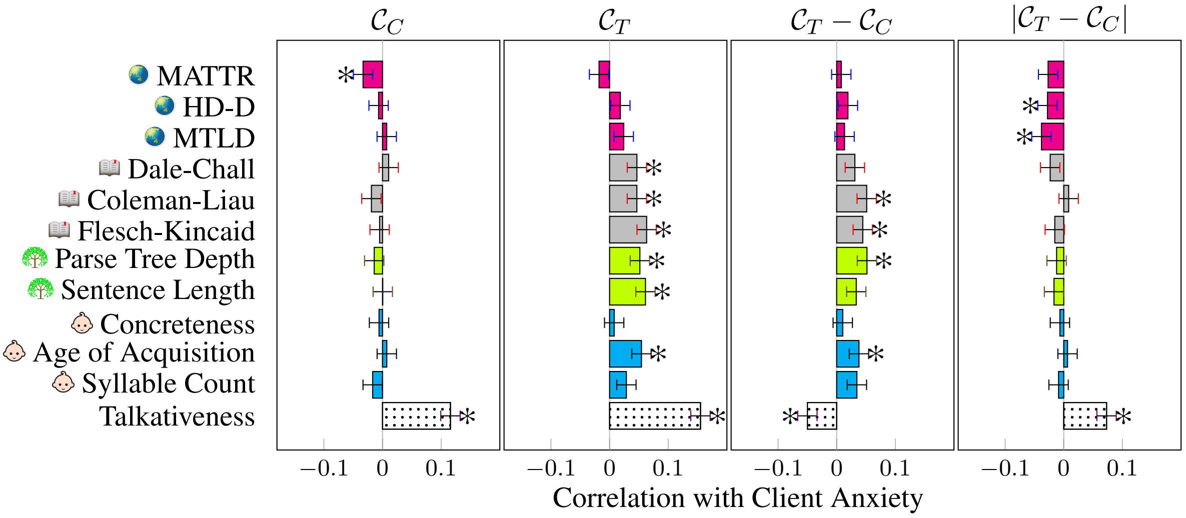

Figure 1: Linguistic complexity measures correlate with client anxiety (✻ indicates significance at $p < 0.001$ for both the exploratory and confirmatory datasets). We show correlations (±99.9% confidence intervals) on the exploratory dataset for language complexity of clients ($\mathcal{C}_C$), therapists ($\mathcal{C}_T$), therapist and client difference ($\mathcal{C}_T - \mathcal{C}_C$), and absolute therapist and client difference ($|\mathcal{C}_T - \mathcal{C}_C|$). Each complexity measure was entered into its own linear mixed model. We group complexity measures into lexical diversity (🌏), syntax (🌳), readability (📖), and prototypicality (👶).

model separately for each of our eleven measures (and talkativeness) and report the normalized correlation coefficient of $\mathcal{C}^m$ on anxiety. A further description of our linear mixed model can be found in the Appendix.

Figure 1 (first and second panels) shows these results for client linguistic complexity $\mathcal{C}_C$ and therapist linguistic complexity $\mathcal{C}_T$. For clients, most linguistic complexity measures had non-significant or slightly negative correlations with anxiety. Moving average type-token ratio (MATTR), which measures the ratio of unique words while accounting for sequence length, was the only significant predictor of anxiety. This correlation was negative, suggesting that clients showed less lexical diversity when they were stressed and providing some evidence that the complexity loss paradox might manifest in language—higher anxiety co-occured with less diverse word choice, a form of linguistic complexity loss. HD-D and MTLD, the two other estimation techniques for lexical diversity, not decrease significantly with higher anxiety. HD-D samples words randomly and is thus unaffected by word order whereas MATTR does account for word order, suggesting that the relationship between decreased word diversity and anxiety might exist in local linguistic structure rather than global word usage; MTLD uses a previously established threshold based on books, whereas MATTR does not use thresholding. These measures, which take varying approaches to estimating lexical diversity, relate differentially to anxiety; we leave investigating this phenomenon's underpinnings as future work.

Therapist language, on the other hand, showed higher reading difficulty, syntactic complexity, and age of acquisition when clients were more anxious, potentially reflecting a therapist's responsiveness to their client's current states. Therapists listen closely to what clients say, and through reviewing survey results, build intuitions on clients' mental states. They also undergo extensive training before being licensed on Talkspace, and so we speculate that when clients are more anxious, therapists are more likely to have detailed and involved discussions with clients, which can involve more complex language due to the sensitive nature of the conversation topics. In addition, both clients and therapists were more verbose (higher talkativeness) when clients were more anxious.

In addition to $\mathcal{C}_C$ and $\mathcal{C}_T$, we also investigate how difference in client and therapist language $\mathcal{C}_T - \mathcal{C}_C$ and similarity between client and therapist language $|\mathcal{C}_T - \mathcal{C}_C|$ correlate with anxiety (Figure 1, third and fourth panels). For $\mathcal{C}_T - \mathcal{C}_C$, therapist language had higher measures of Coleman-Liau, Flesch-Kincaid, parse tree depth, and age of acquisition than client language when clients were more anxious. For $|\mathcal{C}_T - \mathcal{C}_C|$, smaller differences in HD-D and MTLD predicted lower client anxiety, suggesting that therapist and client lexical diversity was more similar when clients were less stressed.

	Standard Deviation (σ)				Range (Δ)			
			$z_C^\sigma \neq z_T^\sigma$?				$z_C^\Delta \neq z_T^\Delta$?	
	z_C^σ	z_T^σ	t	p	z_C^Δ	z_T^Δ	t	p
🌍 MATTR	-0.36	-0.33	-0.87	0.3842	-0.35	-0.29	-2.17	0.0298
🌍 HD-D	-0.3	-0.32	0.53	0.5936	-0.3	-0.28	-0.83	0.4082
🌍 MTLD	-0.36	-0.35	-0.06	0.9561	-0.35	-0.34	-0.17	0.8665
📖 Dale-Chall	-0.46	-0.65	6.43*	<0.0001	-0.45	-0.51	1.91	0.0563
📖 Coleman-Liau	-0.68	-0.74	1.91	0.0558	-0.66	-0.61	-1.84	0.0664
📖 Flesch-Kincaid	-0.46	-0.93	15.68*	<0.0001	-0.47	-0.76	11.33*	<0.0001
🌳 Parse Tree Depth	-0.77	-0.89	4.12*	<0.0001	-0.74	-0.73	-0.48	0.6324
🌳 Sentence Length	-0.44	-0.97	17.69*	<0.0001	-0.45	-0.79	13.54*	<0.0001
😐 Concreteness	-0.49	-0.36	-4.64*	<0.0001	-0.48	-0.32	-6.57*	<0.0001
😐 Age of Acquisition	-0.48	-0.69	6.15*	<0.0001	-0.47	-0.51	1.44	0.1494
😐 Syllable Count	-0.44	-0.44	0.01	0.9913	-0.43	-0.36	-2.21	0.0273
Talkativeness	-0.31	-0.66	13.48*	<0.0001	-0.3	-0.58	11.88*	<0.0001

Table 3: z indicates how much individuals varied linguistic complexity among their own messages compared with a random sample from the population. We show average z for within-individual standard deviation σ and range Δ for clients C and therapists T. * indicates significance at $p < 0.001$ for both exploratory and confirmatory datasets.

5 Individual Variation in Linguistic Complexity Measures

In addition to assessing whether linguistic complexity measures reflect mental health, we explore the extent to which individuals produce consistent values of complexity measures. Was the complexity profile of a given client or therapist stable across their messages, or did it vary over time?

Because our dataset has a large number of individuals and a varying number of samples per individual, traditional analyses for exploring between-individual and within-individual variation (e.g., ANOVA) were inadequate. Therefore, we take an approach that compares within-individual variation with the expected variation from a random sample in the population, while accounting for the varying numbers of conversations per individual.

For a given individual and complexity measure, we first compute that individual's standard deviation σ among their n conversations. Then, we use σ to generate a z-score z^σ by comparing σ with the distribution of standard deviations given by 1,000 random samples of the same size (same n conversations) from the *entire* population. If the distribution of z^σ for all individuals did not significantly differ from $\mathcal{N}(0,1)$—the expected distribution of z-scores if there were no individual differences—then individuals did not have consistent levels of that complexity measure. If the distribution of individual z-scores was significantly more negative than $\mathcal{N}(0,1)$, however, then individuals had more consistent values of that measure than expected and therefore had *unique voices*. We compute z^σ, as well as z^Δ for ranges Δ, for both clients and therapists.

Table 3 shows average z^σ and z^Δ for clients and therapists. All z-distributions skewed negative (in fact, all z-distributions differed from $\mathcal{N}(0,1)$ with $p < 10^{-8}$), indicating that both clients and therapists had significantly consistent linguistic complexity among their own messages compared with random samples from all messages. Now, given the z distributions for clients and therapists, we use a two-tailed t-test to explore whether these distributions differ. As shown in Table 3, standard deviations for six metrics suggested that therapists had more unique voices, four of which were confirmed by the same analysis for range (compared with clients having more unique voices only for concreteness), possibly an indication of therapists' unique styles of therapy.

6 Conclusions

We have studied linguistic complexity in online therapy conversations as it relates to mental health. We found that clients used less lexically diverse language as estimated by MATTR when they were more anxious, supporting prior work that complexity loss due to anxiety may manifest in word diversity (Connely, 1976). In addition, we found that language of therapists also correlated with client anxiety and was generally more consistent than that of clients. Our work shows that analyzing linguistic complexity can identify meaningful patterns in mental health, an important prospect in an era of both increased online communication and mental health illness (Van den Eijnden et al., 2008).

Acknowledgements We thank Derrick Hull and Talkspace for their generous collaborative efforts and access to the Talkspace dataset.

7 Ethical Considerations

The dataset in this paper is of a sensitive nature, and there are several associated ethical considerations. Our study procedures were approved as exempt by the Committee for the Protection of Human Subjects at Dartmouth. All patients and clinicians gave consent for the use of their data in a de-identified, aggregate format and the dataset is not publicly available. All patients were able to opt out at any time by informing their therapist or contacting support. We emphasize that the findings in our paper are specific to this dataset and we make no claims about their generalizability to other contexts. Our study was a non-clinical investigation of the complexity loss paradox in psychology, as opposed to a psychiatric study designed for clinical or practical applications. Finally, the data (text messages) were written in English and therefore we do not claim that our findings generalize to other languages. For these reasons, we advise caution when working in this domain and building upon these results.

References

Concepción L Alados and Michael A Huffman. 2000. Fractal long-range correlations in behavioural sequences of wild chimpanzees: a non-invasive analytical tool for the evaluation of health. *Ethology*, 106(2):105–116.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Nancy JC Andreasen and Bruce Pfohl. 1976. Linguistic analysis of speech in affective disorders. *Archives of General Psychiatry*, 33(11):1361–1367.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46:904–911.

Robert Cabin and Randall Mitchell. 2000. To bonferroni or not to bonferroni: When and how are the questions. *Bulletin of the Ecological Society of America*, 81:246–248.

Meri Coleman and TL Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Dwight Connely. 1976. Some effects of general anxiety and situational stress upon lexical diversity, speaking rate, speaking time, and evaluations of a speaker.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Manuelle Cottin, Andrew JJ MacIntosh, Akiko Kato, Akinori Takahashi, Marion Debin, Thierry Raclot, and Yan Ropert-Coudert. 2014. Corticosterone administration leads to a transient alteration of foraging behaviour and complexity in a diving seabird. *Marine Ecology Progress Series*, 496:249–262.

Michael Covington and Joe McFall. 2010. Cutting the gordian knot: The moving-average type-token ratio (mattr). *Journal of Quantitative Linguistics*, 17:94–100.

Edgar Dale and Jeanne Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27:11–20.

Egdar Dale and Jeanne Chall. 1995. *Readability revisited: The new Dale–Chall readability formula*. Brookline Books, Cambridge, UK.

Karthik Dinakar, Jackie Chen, Henry Lieberman, Rosalind Picard, and Robert Filbin. 2015. Mixed-initiative real-time topic modeling & visualization for crisis counseling. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 417–426.

Marcel Fafchamps and Julien Labonne. 2016. Using split samples to improve inference about causal effects. Working Paper 21842, National Bureau of Economic Research.

Andrzej Galecki and Tomasz Burzykowski. 2013. *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. Springer Publishing Company, Incorporated.

Amy L Goldberger. 1997. Fractal variability versus pathologic periodicity: complexity loss and stereotypy in disease. *Perspectives in Biology and Medicine*, 40(4).

Ary L Goldberger, C-K Peng, and Lewis A Lipsitz. 2002. What is physiologic complexity and how does it change with aging and disease? *Neurobiology of aging*, 23(1):23–26.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/.

Marijke Höweler. 1972. Diversity of word usage as a stress indicator in an interview situation. *Journal of Psycholinguistic Research*, 1(3):243–248.

Christine Howes, Matthew Purver, and Rose McCabe. 2014. Linguistic indicators of severity and progress in online text-based therapy for depression. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 7–16, Baltimore, Maryland, USA. Association for Computational Linguistics.

Thomas D Hull, Philippa Connolly, Kush Mahan, and Katie Yang. 2018. The treatment effectiveness of asynchronous text therapy for depression and anxiety: A longitudinal cohort study. *Talkspace Research*.

J. Peter Kincaid, Robert P. Fishburne, Richard Lawrence Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44:978–990.

Fei-Tzin Lee, Derrick Hull, Jacob Levine, Bonnie Ray, and Kathy McKeown. 2019. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 12–23, Minneapolis, Minnesota. Association for Computational Linguistics.

OP Malhotra and S Roy Salam. 1991. Cyclic oscillations of leucocyte counts in chronic myeloid leukaemia. *Postgraduate medical journal*, 67(783):87–89.

Philip Mccarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42:381–92.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Philip M McCarthy and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488.

Gerard Mckee, D.D. Malvern, and Brian Richards. 2000. Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, 15:323–337.

Jones G Parker, Jesse D Marshall, Biafra Ahanonu, Yu-Wei Wu, Tony Hyun Kim, Benjamin F Grewe, Yanping Zhang, Jin Zhong Li, Jun B Ding, Michael D Ehlers, et al. 2018. Diametric neural ensemble dynamics in parkinsonian and dyskinetic states. *Nature*, 557(7704):177–182.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language. use: our words, our selves. *Annual Reviews of Psychology*, 54.

Verónica Pérez-Rosas, Xuetong Sun, Christy Li, Yuchen Wang, Kenneth Resnicow, and Rada Mihalcea. 2018. Analyzing the quality of counseling conversations: the tell-tale signs of high-quality counseling. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, and Christopher Brew. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5:BII–S9042.

Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Lowe. 2006. A brief measure for assessing generalized anxiety disorder. *JAMA Internal Medicine*, 166(10).

Talkspace. 2020. Talkspace - #1 rated online therapy, 1 million+ users. https://www.talkspace.com/, accessed Oct 1, 2020.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Brian MacWhinney, and Chris Dyer. 2016. Learning the curriculum with Bayesian optimization for task-specific word representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 130–139, Berlin, Germany. Association for Computational Linguistics.

Regina JJM Van den Eijnden, Gert-Jan Meerkerk, Ad A Vermulst, Renske Spijkerman, and Rutger CME Engels. 2008. Online communication, compulsive internet use, and psychosocial well-being among adolescents: A longitudinal study. *Developmental psychology*, 44(3):655.

Robert S Wachal and Otfried Spreen. 1973. Some measures of lexical diversity in aphasic and normal language performance. *Language and Speech*, 16(2):169–181.

Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019. Finding your voice: The linguistic development of mental health counselors. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 936–947, Florence, Italy. Association for Computational Linguistics.

8 Appendix

§8.1 defines and explains our linguistic complexity measures in further detail. To supplement §4, §8.2 details our linear mixed model.

8.1 Definitions of Complexity Measures

Here, we describe in detail the linguistic complexity measures we used, which span lexical diversity (🌏), syntactic simplicity (🌳), readability (📖), and prototypicality (👶).

1. 🌏 Type-Token Ratio (TTR): #types / #tokens. Because TTR decreases for longer texts, we use the moving average type-token ratio (MATTR) (Covington and McFall, 2010)—for a given sequence of tokens, we slide a window of size $W = 50$ over all tokens with a stride of $s = 1$, compute lexical richness for each of the windows, and output the average.

2. 🌏 HD-D (vocd-D): McCarthy and Jarvis (2007) found that output $\mathcal{D}$ of vocd-D (Mckee et al., 2000), which estimates the fit of TTRs for text samples of different length, is merely a complex approximation ($R = 0.971$) of a hypergeometric distribution, which they use in an index called HD-D.[1] HD-D measures the mean contribution that each type makes to the TTR of all possible combinations of a samples of size 35-50, and higher HD-D indicates greater lexical diversity

3. 🌏 Measure of Textual Lexical Diversity (MTLD): this more complicated measure of lexical diversity measures the mean length of word strings that maintain a criterion level of lexical variation. See McCarthy (2005) for details.

4. 🌳 Parse tree depth: dependency parse tree depth using `spaCy` (Honnibal and Montani, 2017).

5. 🌳 Sentence length: number of words in a sentence.

6. 📖 Dale-Chall readability score (Dale and Chall, 1948, 1995): DCRS $= 0.1579($DWR $\cdot\ 100) +$ 0.0496WPS, where DWR is the ratio of difficult words[2] and WPS is the average words per sentence. Texts with higher DCRS are supposed to be more challenging to read.

7. 📖 Coleman-Liau index (Coleman and Liau, 1975): CLI $= 0.0588L - 0.296S - 15.8$, where L is the average number of letters per 100 words and S is the average number of sentences per 100 words. CLI aims to approximate the U.S. grade level thought necessary to comprehend the text.

8. 📖 Flesch-Kincaid grade level (Kincaid et al., 1975): FKGL $= 0.39$WPS $- 11.8$SPW $- 15.59$, where WPS is the average words per sentence and SPW is the average syllables per word. Higher scores indicate material that is more challenging to read.

9. 👶 Age of acquisition (AoA): the average AoA of words was extracted from a database of crowdsourced ratings of over 30 thousand words (Kuperman et al., 2012). For instance, *potty* has an AoA of 2.28, and *blasphemous* has an AoA of 11.25.

10. 👶 Concreteness: averaged word-level concreteness ratings on the scale from 1–5 (1 is most abstract, and 5 is most concrete) for 40 thousand English lemmas (Brysbaert et al., 2014). For instance, *spirituality* is rated 1.07, and *scarf* is rated 4.97.

11. 👶 Syllable count: average number of syllables per word, as computed using `pyphen`: `https://pyphen.org/`

12. Talkativeness: number of alphanumeric tokens for either client or therapist in a conversation (all messages in one week period before a survey).

8.2 Linear Mixed Model Analysis

As the anxiety of clients can correlate with many variables, we use a linear mixed model (Galecki and Burzykowski, 2013) (sometimes called multilevel or hierarchical models), which is a regression model that accounts for both fixed effects (variation that is explained by independent variables of interest) and random effects (variation that is not explained by independent variables of interest). In this subsection, we show the expressions for the linear mixed models we use in §4. For each linguistic complexity measure m, our fixed effects include client linguistic complexity $\mathcal{C}^m_C$, therapist linguistic complexity $\mathcal{C}^m_T$, time (weeks in therapy) t, client

[1] See `https://textinspector.com/help/lexical-diversity/` for McCarthy's recommendation on vocd-D vs HD-D.

[2] Words not on a list of 3,000 familiar words at `https://www.readabilityformulas.com/articles/dale-chall-readability-word-list.php`

age a_C, client gender g_C, and client education e_C, and our random effects include clients with respect to time $(1 + t|C)$, and therapists $(1|T)$.

For computing correlation between client anxiety and a linguistic complexity variable of interest $\mathcal{C}' \in \{\mathcal{C}^m_C, \mathcal{C}^m_T, (\mathcal{C}^m_C - \mathcal{C}^m_T), |\mathcal{C}^m_C - \mathcal{C}^m_T|\}$ For computing correlations between linguistic complexity and client anxiety (Figure 1), we use

$$\begin{aligned} \text{anxiety} \sim \mathcal{C}' + t + a_C + g_C + e_C \\ + (1 + t|C) + (1|T) \ . \end{aligned} \tag{1}$$