Analyzing the Quality of Counseling Conversations: the Tell-Tale Signs of High-quality Counseling

Verónica Pérez-Rosas[1], **Xuetong Sun**[1], **Christy Li**[1], **Yuchen Wang**[1]
Kenneth Resnicow[2] **and Rada Mihalcea**[1]
[1]Computer Science and Engineering, [2] School of Public Health
University of Michigan
{vrncapr, xuetong, christyq, lucasyc,kresnic,mihalcea}@umich.edu

Abstract

Behavioral and mental health are pressing issues worldwide. Counseling is emerging as a core treatment for a variety of mental and behavioral health disorders. Seeking to improve the understanding of counseling practice, researchers have started to explore Natural Language Processing approaches to analyze the nature of counseling interactions by studying aspects such as mirroring, empathy, and reflective listening. A challenging aspect of this task is the lack of psychotherapy corpora. In this paper, we introduce a new dataset of high-quality and low-quality counseling conversations collected from public web sources. We present a detailed description of the dataset collection process, including preprocessing, transcription, and the annotation of two counseling micro-skills: reflective listening and questions. We show that the obtained dataset can be used to build text-based classifiers able to predict the overall quality of a counseling conversation and provide insights into the linguistic differences between low-quality and high-quality counseling.

Keywords: conversation analysis, counseling, mental health

1. Introduction

Mental and behavioral disorders, such as substance abuse, are top on the list of the most costly and prevalent conditions worldwide.[1] Particularly in the US, a recent survey on public health reported that in 2014 3.3% of all adults had co-occurring mental illness and substance abuse disorders.[2]

As behavioral counseling has been been shown to be an effective treatment method for these conditions, the number of people seeking counseling services is increasing (Chava, 2014). Despite its potential benefits, such as combating addiction and providing broader disease prevention and management, the mechanisms behind successful behavioral counseling have not been fully elucidated (Moyers et al., 2009).

Specific counseling skills have shown to increase the likelihood of positive health outcomes (Gaume et al., 2009; Vader et al., 2010). Regardless of the counseling method, counselors follow general principles, such as supporting autonomy, expressing empathy, centering on the patient and engaging patients using specific skills such as reflective listening (Charles et al., 1997; Harting et al., 2004). In contrast, using a more directing style – characterized by counselors providing instruction and advice, and patients obeying, adhering and complying (Miller and Rollnick, 2013) – is usually avoided.

The guidelines described above can be used to differentiate between low and high quality counseling. Thus, in a broad classification, psychotherapy conversations where counselors follow preferred practices can be considered as high-quality (or guideline adhering) counseling, whereas those conversations where they do not can be regarded as low-quality counseling (or guideline non-adhering).

Following this idea, our paper analyzes counseling conversations with the final goal of distinguishing between low and high quality counseling. In particular, we focus our analysis on counseling conducted using Motivational Interviewing (MI), a well-established evidence-based counseling style for treating addiction and other behaviors (Moyers et al., 2009; Catley et al., 2012; Apodaca et al., 2014).

Our work makes two main contributions. First, we introduce a new dataset of counseling conversations collected from public web sources. With this dataset, we seek to address the problem of lack of psychotherapy corpora for NLP applications, as most of current psychotherapy corpora have important constrains regarding their public accessibility due to ethical and privacy concerns. Second, we show that the collected dataset can be used to build text-based classifiers able to predict the overall quality of a counseling conversation and provide insights into the tell-tale signs of high-quality counseling.

2. Related Work

While clinical mental health counseling has been found useful in the treatment of public health issues, evaluating its quality remains a problem. This is mainly because most studies on clinical psychology have been limited by the need for human-based evaluation and by small sample sizes.

Computational approaches for the analysis of counseling interactions have focused on two main lines of work.

First, seeking to develop tools for the automatic evaluation of counseling practice, several linguistic based approaches have been proposed to aid the automatic identification of counselor and client behaviors that are correlated to successful interventions (Klonek et al., 2015). (Can et al., 2012) used n-grams, similarity features between coun-

[1]Word Health Report 2001, http://www.who.int/whr/2001/media_centre/en/

[2]The State of Mental Health in America https://www.samhsa.gov/disorders/substance-use

selor and client speech, and dialog meta-features to automatically detect and code counselors' reflective listening. A method based on labeled topic models is presented in (Atkins et al., 2012; Atkins et al., 2014), where authors focus on automatically identifying conversation topics that relate to counselor behaviors such as reflective listening, questions, support, and empathy. Methods that combine acoustic and linguistic datastreams have also been proposed to evaluate the quality of counseling interactions. (Xiao et al., 2014) presented a study on the automatic evaluation of counselor empathy based on analyzing correlations between prosody patterns and empathy showed by the therapist during counseling interactions.

Second, aiming to improve the understanding of counseling interactions, researchers have started to explore NLP approaches to study aspects such as language mirroring, empathy, and reflective listening. (Tanana et al., 2015) addressed the identification of counselor's statements that discuss client's change talk using recursive neural networks to model sequences of counselor and client verbal exchanges. (Lord et al., 2015) analyzed the language style synchrony between therapists and clients during MI encounters. Their approach relies on the psycholinguistic categories from the Linguistic Inquiry and Word Count lexicon to measure the degree to which the counselor language matches the client language. More recently, (Althoff et al., 2016) explored language style and symmetry in counseling interactions by analyzing a large sample of text-message-based counseling. Their main findings suggest that counselors who are more successful act with more control in the conversations and show lower levels of verbal coordination (mirroring) than their less successful counterparts.

Furthermore, there are ongoing efforts on creating annotated resources that facilitate NLP advances in the analysis of clinical text in applications such as automatic annotation of pathology reports and oncology reports as well as data from biomedical journals (Roberts et al., 2007; Albright et al., 2013; Verspoor et al., 2012). Despite this efforts, to our knowledge, there are only few psychotherapy corpora available. One of them is the "Alexander Street Press", [3] which is a large collection of transcripts and video recordings of therapy sessions on different subjects such as anxiety, depression, family conflicts, and others. There are also other psychology datasets available under limited access from the National Institute of Mental Health (NIMH).[4]

In this paper, we present the development of a counseling conversations dataset that can be used to implement data-driven methods for the automatic evaluation of counseling quality. We specifically focus on the overall conversation quality, with the final goal of providing linguistic cues associated with high-quality counseling.

[3]http://alexanderstreet.com/products/counseling-and-psychotherapy-transcripts-series

[4]http://psychiatry.yale.edu/pdc/resources/datasets.aspx

3. A Dataset of Low and High Quality Counseling

3.1. Collecting Counseling Conversations from the Web

We started by identifying video clips containing brief counseling interactions conducted using Motivational Interviewing (MI) from publicly available video-sharing sources such as YouTube and Vimeo. Keywords used to search for these videos include "motivational interviewing", "MI counseling", "effective MI", "good MI", "MI counseling demonstration", "role play MI" for the high-quality category, and "ineffective MI", "bad MI", "bad counseling", "how not to do MI", "the bad counselor" for the low-quality category. To select the videos, we used the following guidelines: the video should include only two participants, i.e., counselor and client; the video should include minimal interruptions, such as background narrative, music, or animation; the session should address a behavior change e.g. smoking cessation, drinking; and finally, the counselor-client interaction should last at least 3 minutes.

The obtained recordings consist mainly of MI counseling demonstrations from MI training services and students' MI role-play practice from undergraduate-level psychology courses. The sessions address various health topics including smoking cessation, alcohol consumption, substance abuse, weight management, mental disorders, and medication adherence and portray several practice settings such as private practice, school counseling, and pharmacy counseling among others.

After collecting our initial pool of videos, we conduct a second filtering step to verify that the counseling was conducted using MI and that the video caption matched the video content, i.e., portray either a high-quality or a low-quality counseling interaction. To evaluate MI use (or the lack of it) we followed the guidelines in MI literature (Miller and Rollnick, 2013). The criteria to label a counseling interaction as either low or high quality is as follows: during high-quality counseling, the conversation should present, to some extent, reflective listening, questions, as well as collaboration and support. In contrast, the low-quality counseling should show a predominant directive style, which includes confrontation, advising without permission, and lack of listening.

The final video set includes 151 counseling conversations. From this, 72 video clips were labeled as high-quality counseling and the remaining 79 as low-quality counseling. The length of the conversations varies from 5-20 minutes. Table 1 shows transcript excerpts corresponding to high-quality and low-quality counseling conversations in the dataset.

Preprocessing. All the videos are first converted into standard mp4 format and then preprocessed to address issues frequently present in shared video content such as introductory titles, animations, and narratives. In most cases these interruptions appeared only at the beginning of the video so we manually trimmed that portion of the video until the counselor-patient interaction started. This process can also be optimized using automatic methods such as optical character and facial recognition, however, we opted for a manual approach in order to obtain accurate examples

HIGH-QUALITY COUNSELING	LOW-QUALITY COUNSELING
T: Hi miss NAME my name is NAME. I'm a social worker at the Family Health Center *C*: Dr. NAME asked me if I would spend some time with you today *T*: I'm really glad that you here. I'm just curious as to why he would send you to me *C*: well I came to see Dr. Steele last week because of increasing stress and anxiety. That's kind of getting the best of me and in the course of my appointment with him he was asking how I was dealing with that stress and I mentioned that my one or two glasses of wine a few nights a week is turning into more frequent. *T*: yeah so he he actually see me because he went to him for increased stress and he's concerned that your alcohol consumption may be a part of that increase and prior to prescribing you anything you want to make sure that you had someone to talk to about that	*T*: Okay, so I wrote a prescription for an antibiotic for NAME that should help with the ear infection but in looking through this chart, I mean, it seems like he's had six or seven of these just in the past year or so that's really a big problem *C*: Yeah it's pretty stressful for both of us. It gets really upset *T*: Well, one of the primary risk factors for multiple ear infections and kids is actually smoke exposure. Are you smoking? *C*: Yeah. I, yeah, I do smoke but I don't smoke around him. I try really hard not to smoke around him *T*: Well, the fact that he's having these ear infections is indicating to me that he is being exposed to smoke and so what can you tell me about that?

Table 1: Transcript excerpts corresponding to high-quality and low-quality counseling conversations

Code	Count	Verbal examples
Question	1122	What do you think it would take to change your mind about participating in physical activity?
Reflection	813	It sounds like you're concerned by your weight and you want to start to make positive changes.

Table 2: Frequency counts and verbal examples of Questions and Reflections in the dataset

of counselor-patient interactions.

Transcription. In order to transcribe the video clips, we adopt a semi-automatic approach. First, we use the YouTube automatic captioning to obtain the conversation transcript and then we manually labeled the conversation turns as either counselor or patient speech.

3.2. Annotation of Counseling Skills

Seeking to evaluate the counseling interaction between counselors and patients, we decided to annotate two core interview micro-skills in counseling practice: reflective listening and questions (Tollison et al., 2008). These two micro-skills are assessed using the Motivational Interviewing Treatment Integrity (MITI) coding scheme version 4 (Moyers et al., 2016), which is the current gold standard for MI fidelity evaluation. Hence, all the video clips in our dataset are manually annotated to identify questioning (Questions) and reflective listening statements (Reflections).

In order to conduct the annotation, two undergraduate students were trained in the use of the MITI 4.0. During this training, the annotators learned how to parse the counselor-patient interaction (i.e., deciding which portion of the conversation shows the given behavior), practiced the correct assignment of behavior codes, and conducted team coding on sample sessions.

The 151 sessions were randomly distributed among the two annotators. During the coding process, the annotators used both the audio recording and the transcript. The annotation was conducted at conversation turn-level using Nvivo,[5] an annotation and quantitative analysis suite that allows selecting text in the transcript and labeling it with a given code, e.g., reflection or question.

In order to verify the reliability of the annotations, we calculated the inter-annotator agreement in a sample of 20 counseling conversations, with even distribution for the Low-quality and High-quality categories. The intra-class correlation scores for both Questions and Reflections codes are 0.96 and 0.94, respectively, thus showing good levels of agreement between the two annotators.

4. Discriminating Between High-quality and Low-quality Counseling

4.1. Analysis of Counseling Conversations

We start by exploring linguistic differences between the counseling interactions to get insights into the mechanisms of high-quality counseling. Our analyses are based on the semantic word classes from the LIWC lexicon and the semantic word-class scoring by (Mihalcea and Pulman, 2009). Table 3 shows the top classes for both, low and high quality counseling.

The results show interesting differences between the two types of conversations. While high-quality counseling focus on aspects related to encouragement and reflective listening such as family, positive feeling, feelings, and hearing, low-quality counseling shows a more directive lan-

[5]http://www.qsrinternational.com/what-is-nvivo

High quality Counseling			Low quality Counseling		
Class	Score	Sample words	Class	Score	Sample words
Family	2.04	Mom, wife, parents, husband	Self	1.33	I, we, me, my, our
Feel	1.85	Feel, pain, feeling, sense	Negate	1.31	Not, don't, no, can't, without
Posfeel	1.52	Like, care, enjoy, glad	Inhibition	1.81	quit, stop, control, avoid
Anxiety	1.38	Afraid, worried, overwhelmed	Time	1.06	Now, start, today, before
Optimism	1.31	Ready, hope, confidence, determined	Present	1.06	Know, do, need, want
Hear	1.25	Sounds, heard, talking, said	Pronoun	1.04	You, I, it, your, we

Table 3: Results from LIWC word class analysis. Top ranked semantic classes associated to low and high quality counseling are shown.

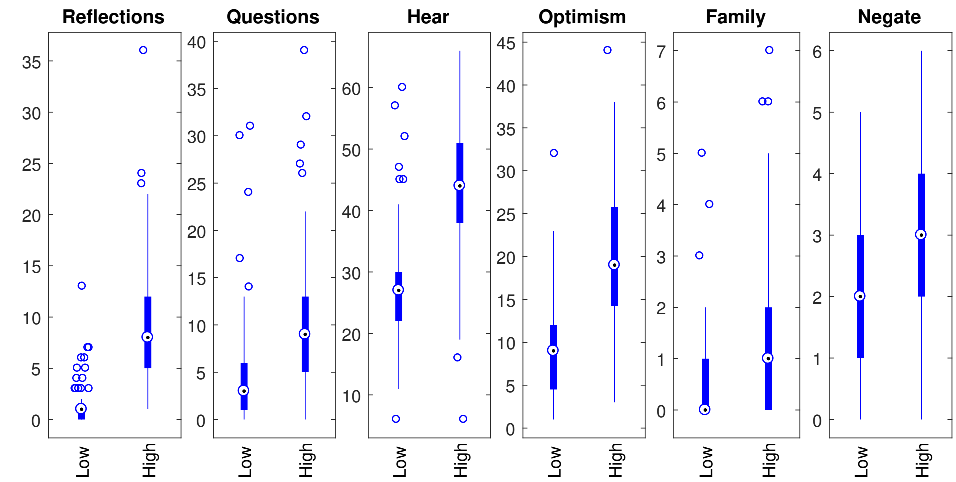

Figure 1: Distribution of reflections, questions, listening (hear), optimism, family and negation (negate) word classes in high and low quality counseling

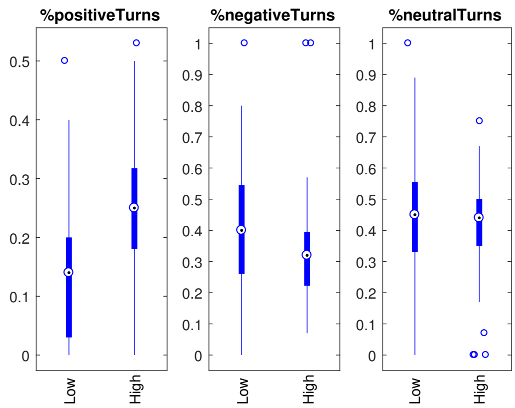

Figure 2: Distribution of positive, negative, and neutral counselor turns in high and low quality counseling

guage by using more self-references, inhibition and negation words.

To further analyze these trends, we plot the distribution of reflections, questions, and lexicon word-classes over the two types of counseling conversations. Results shown in Figure 1 show important differences between the low and high quality groups, thus suggesting that they are potentially good predictors of counseling quality.

In addition, we analyzed the sentiment expressed by the counselor during the encounters as a potential predictor of counseling quality. This could provide information on whether counselors focus on positive or negative aspects of the client communication, and how this relates to the conversation quality, i.e., low or high. Thus, we analyze the sentiment expressed by counselors during each turn in the conversation. Given the effort required to manually annotated the sentiment in each conversation, we opted for using an automatic off-the-shelf sentiment classifier from the Stanford Core NLP package (Manning et al., 2014). We obtain a sentiment score for each counselor turn, scored from very negative to very positive, and calculate the percentage of positive, negative, and neutral turns during the conversation. Figure 2 shows the sentiment distribution over the high-quality and low-quality conversations. The box plots in the figure suggest differences between the low and high quality groups, particularly for positive sentiment. In order to look more closely into the positive sentiment trend during the counseling encounters, we plot the distribution of positive turns by the counselor across the low and high quality counseling conversation. The plot in Figure 3 shows that counselors increasingly focus on positive aspects of the client expressions, particularly during high-quality conversations.

4.2. Distinguishing Between High and Low Quality Counseling

In this section, we explore the use of linguistic cues to build a computational model that predicts the overall quality of

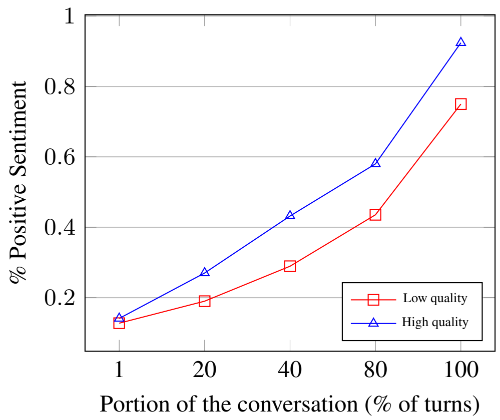

Figure 3: Positive sentiment across five equal segments of the conversation duration.

the counseling conversation
The feature set consists of the cues identified during our exploratory analyses as potential indicators of counseling quality, as well as additional text features used during standard NLP feature extraction such as ngrams. The features are extracted from the transcripts of the counseling conversations. During our experiments, we first explore the predictive power of each cue separately, followed by an integrated model that attempts to combine all the linguistic cues to improve the prediction of counseling quality. The different features are as follows:

***N-grams*:** These features represent the language used by the counseling-conversation participants and include all the unique words and word-pairs present in the transcript. We extract a vector containing the frequencies of each word and word pair present in the transcript.

***Semantic information*:** We use categories from the LIWC (Tausczik and Pennebaker, 2010), Opinion Finder (Wilson et al., 2005) and the Wordnet Affect (Strapparava and Valitutti, 2004) lexicons to derive features that identify identifying words as belonging to certain semantic categories that are potential markers of the conversation quality.

Metafeatures: We also extract a set of metafeatures that describe the conversation interaction, including the number of counselor turns, client turns, average words during client and counselor turns, and the ratio of counselor and client words in each turn.

Sentiment: These features are designed to capture the sentiment trend in the counselor responses during the conversation. To derive this features, we first obtain the sentiment expressed by counselors during each turn, scored from very negative to very positive $(--, -, 0, +, ++)$ using the sentiment analysis classifier from Stanford Core NLP, and then obtained a set of descriptors that capture the sentiment trend. The set includes the percentage of positive, negative, and neutral turns during the conversation, the number of times the sentiment changes during the conversation, as well

Feature set	Acc.	F-score High	F-score Low
Baseline	52.31%		
Ngrams	82.78%	0.82	0.83
Lexicons	72.84%	0.71	0.73
Metafeatures	76.15%	0.77	0.74
MITI Behav	83.44%	0.83	0.83
Sentiment	70.86%	0.69	0.81
All features	**87.41%**	0.87	0.87

Table 4: Overall prediction results and F-scores for high-quality and low-quality counseling conversations using several linguistic feature sets.

Feature set	Acc.
All features	**87.41%**
– Ngrams	83.44%
– Lexicons	85.43%
– Sentiment	86.10%
– Metafeatures	87.41%
– MITI Behav	87.41%

Table 5: Feature ablation study.

as counts of sequences increasing and decreasing sentiment intensity i.e., $-+$, $-++$, $-+$, $+--$.

MITI behaviors This set includes the number of reflections and questions by the counselor during the conversation as well as the ratio of reflections to questions. The counts are derived from the turn-level annotations described in section 3.2.

We conduct several experiments to discriminate between low-quality and high-quality encounters. During our experiments, the evaluations are done at conversation level. The classifiers are built using the Support Vector Machine algorithm[6] and the different sets of linguistic features. We perform leave-one-out cross-validation in all our experiments and we use the majority class baseline as a reference value.
Results shown in Table 4 show that all the feature sets perform above the baseline, with the MITI behaviors being the best performing features, followed by the n-grams features. We also observe that the combination of all feature sets provides the best performance.
Seeking to explore the role played by the different feature sets, we conduct an ablation study, where we remove one feature set at the time from the best performing model i.e., "all features". As observed in Table 5, the ngrams features contribute the most to the final model, followed by the lexicon features. Interestingly, the results show that the combination of n-grams and lexicons offer similar performance as the MITI behaviors features. These results are encouraging as they suggest that standard linguistic features can achieve similar performance as manually coded features (MITI behaviors) while evaluating the overall quality of counseling conversations.

[6] As implemented in the Weka library.

5. Conclusion and Future Work

In this paper, we introduced a new dataset of low-quality and high-quality counseling conversations that were collected from public sources. Through several classification experiments, we showed that such a dataset can be used to build accurate classification models able to discriminate between low-quality and high-quality counseling, with accuracy figures up to 87%.

Furthermore, we showed that standard NLP features can provide performance similar to manually coded features for this task.

We also provided insights into the linguistics markers of high-quality counseling and showed that it is characterized by positive and encouraging language.

6. References

Albright, D., Lanfranchi, A., Fredriksen, A., Styler, W. F., Warner, C., Hwang, J. D., Choi, J. D., Dligach, D., Nielsen, R. D., Martin, J., et al. (2013). Towards comprehensive syntactic and semantic annotations of the clinical narrative.

Althoff, T., Clark, K., and Leskovec, J. (2016). Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*.

Apodaca, T. R., Borsari, B., Jackson, K. M., Magill, M., Longabaugh, R., Mastroleo, N. R., and Barnett, N. P. (2014). Sustain talk predicts poorer outcomes among mandated college student drinkers receiving a brief motivational intervention. *Psychology of Addictive Behaviors*, 28(3):631.

Atkins, D. C., Rubin, T. N., Steyvers, M., Doeden, M. A., Baucom, B. R., and Christensen, A. (2012). Topic models: A novel method for modeling couple and family text data. *Journal of family psychology*, 26(5):816.

Atkins, D. C., Steyvers, M., Imel, Z. E., and Smyth, P. (2014). Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9(1):49.

Can, D., Georgiou, P. G., Atkins, D. C., and Narayanan, S. S. (2012). A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. In *INTERSPEECH*, pages 2254–2257. ISCA.

Catley, D., Harris, K. J., Goggin, K., Richter, K., Williams, K., Patten, C., Resnicow, K., Ellerbeck, E., Bradley-Ewing, A., Malomo, D., et al. (2012). Motivational interviewing for encouraging quit attempts among unmotivated smokers: study protocol of a randomized, controlled, efficacy trial. *BMC public health*, 12(1):456.

Charles, C., Gafni, A., and Whelan, T. (1997). Shared decision-making in the medical encounter: What does it mean? (or it takes at least two to tango). *Social Science & Medicine*, 44(5):681 – 692.

Chava, Z. (2014). Expenditures for mental health among adults, ages 18-64, 2009-2011: Estimates for the u.s. civilian noninstitutionalized population. *Agency for Healthcare Research and Quality, Rockville, MD*, Statistical Brief #454.

Gaume, J., Gmel, G., Faouzi, M., and Daeppen, J.-B. (2009). Counselor skill influences outcomes of brief motivational interventions. *Journal of Substance Abuse Treatment*, 37(2):151 – 159.

Harting, J., van Assema, P., van der Molen, H. T., Ambergen, T., and de Vries, N. K. (2004). Quality assessment of health counseling: performance of health advisors in cardiovascular prevention. *Patient Education and Counseling*, 54(1):107 – 118.

Klonek, F. E., Quera, V., and Kauffeld, S. (2015). Coding interactions in motivational interviewing with computer-software: What are the advantages for process researchers? *Computers in Human Behavior*, 44:284–292.

Lord, S. P., Sheng, E., Imel, Z. E., Baer, J., and Atkins, D. C. (2015). More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior therapy*, 46(3):296–303.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Mihalcea, R. and Pulman, S. (2009). Linguistic ethnography: Identifying dominant word classes in text. In *Computational Linguistics and Intelligent Text Processing*, pages 594–602. Springer.

Miller, W. R. and Rollnick, S. (2013). *Motivational interviewing: Helping people change, Third edition*. The Guilford Press.

Moyers, T. B., Martin, T., Houck, J. M., Christopher, P. J., and Tonigan, J. S. (2009). From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. *Journal of consulting and clinical psychology*, 77(6):1113.

Moyers, T. B., Rowell, L. N., Manuel, J. K., Ernst, D., and Houck, J. M. (2016). The motivational interviewing treatment integrity code (miti 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65(Supplement C):36 – 42. Motivational Interviewing in Substance Use Treatment.

Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J., Roberts, I., Setzer, A., Tapuria, A., and Wheeldin, B. (2007). The CLEF corpus: semantic annotation of clinical text. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 625–629.

Strapparava, C. and Valitutti, A. (2004). WordNet-Affect: an affective extension of WordNet. In *Proceedings of LREC*, volume 4, pages 1083–1086.

Tanana, M., Hallgren, K., Imel, Z., Atkins, D., Smyth, P., and Srikumar, V. (2015). Recursive Neural Networks for Coding Therapist and Patient Behavior in Motivational Interviewing. *2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 71–79.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psy-*

chology, 29(1):24–54.

Tollison, S. J., Lee, C. M., Neighbors, C., Neil, T. A., Olson, N. D., and Larimer, M. E. (2008). Questions and reflections: The use of motivational interviewing microskills in a peer-led brief alcohol intervention for college students. *Behavior Therapy*, 39(2):183 – 194.

Vader, A. M., Walters, S. T., Prabhu, G. C., Houck, J. M., and Field, C. A. (2010). The language of motivational interviewing and feedback: counselor language, client language, and client drinking outcomes. *Psychology of Addictive Behaviors*, 24(2):190.

Verspoor, K., Cohen, K. B., Lanfranchi, A., Warner, C., Johnson, H. L., Roeder, C., Choi, J. D., Funk, C., Malenkiy, Y., Eckert, M., et al. (2012). A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC bioinformatics*, 13(1):1.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005). Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, pages 34–35. Association for Computational Linguistics.

Xiao, B., Bone, D., Van Segbroeck, M., Imel, Z. E., Atkins, D. C., Georgiou, P. G., and Narayanan, S. S. (2014). Modeling therapist empathy through prosody in drug addiction counseling. In *Fifteenth Annual Conference of the International Speech Communication Association*.