Analyzing the Language of Therapist Empathy in Motivational Interview based Psychotherapy

Bo Xiao*, Dogan Can†, Panayiotis G. Georgiou*, David Atkins‡ and Shrikanth S. Narayanan*†
* Department of Electrical Engineering, † Department of Computer Science,
University of Southern California, Los Angeles, CA, U.S.A.
‡ Department of Psychiatry & Behavioral Sciences, University of Washington, Seattle, WA, U.S.A.
*† http://sail.usc.edu/people.php, ‡ datkins@u.washington.edu

***Abstract*—Empathy is an important aspect of social communication, especially in medical and psychotherapy applications. Measures of empathy can offer insights into the quality of therapy. We use an N-gram language model based maximum likelihood strategy to classify empathic versus non-empathic utterances and report the precision and recall of classification for various parameters. High recall is obtained with unigram while bigram features achieved the highest F1-score. Based on the utterance level models, a group of lexical features are extracted at the therapy session level. The effectiveness of these features in modeling session level annotator perceptions of empathy is evaluated through correlation with expert-coded session level empathy scores. Our combined feature set achieved a correlation of 0.56 between predicted and expert-coded empathy scores. Results also suggest that the longer term empathy perception process may be more related to isolated empathic salient events.**

I. Introduction

Empathy is a natural human ability that is studied across disciplines including psychology, neuroscience and social science [1]. Merriam-Webster dictionary defines empathy as "the action of understanding, being aware of, being sensitive to, and vicariously experiencing the feelings, thoughts, and experience of another of either the past or present without having the feelings, thoughts, and experience fully communicated in an objectively explicit manner". In general, empathy stands for the mental ability of feeling for, and taking the perspective of, others.

In social interactions, when empathy is expressed through verbal and non-verbal behaviors, the other party would feel acknowledged, resulting in better and efficient communication. Therefore, showing empathy is deemed to be an important skill and often related to better performance in domains centered on human interaction, such as medical care and psychotherapy [2], [3].

As one form of psychotherapy, Motivational Interview (MI) emphasizes the client's own will of making a change, where the therapist should try to understand the client and facilitate this change, instead of dictating what the client should do. Hence empathy is one of the quality indexes of therapist in MI. Conventionally, empathy is measured by observing audio or audio-visual recordings following expert designed coding manuals. Due to the abstract nature of empathy, coders must be trained to ensure reliability. Training and viewing the recordings require extensive time, and the coding process is difficult to scale up. Researchers are seeking computational techniques to automate this process and provide tools that facilitate their analysis, and multimodal Behavior Signal Processing (BSP) approaches offer promising avenues to address this problem [4], [5]. In addition, BSP aims to not only aid, but also transform observational practice through insights and increased observational capabilities, eg. [6].

As special cases of human interaction, medical care and psychotherapy dialogs are usually more structured, where the conversation follows certain implicit protocol and usually targets specific diagnostic and informational goals. Manifestation of care-provider's empathy is embedded in their communication cues and patterns. One of the key sources of such cues is the language use. This paper focuses on computationally analyzing empathy behavior expressed in spoken language information. There is promising support for this line of work. In [3] the domain experts studied empathy behavior exemplified through transcripts of the conversation. Moreover, language modeling towards classification of a group of abstract behaviors (e.g., acceptance, blame, humor, etc.) in distressed couple interactions has been shown to be effective, even with Automatic Speech Recognition (ASR) derived lexical features [4]. This motivates us to study computationally the relation of empathy expression and the corresponding language use.

In this paper we utilize two datasets of MI based psychotherapy sessions. In the first set empathic language is annotated at the utterance level, from which we learn empathic and non-empathic language models. Precision and recall in experiments of classifying utterances into empathic or non-empathic classes are reported. The second set of sessions are given a session level score of therapist empathy. We use the language model learned on the first set to extract a group of session level lexical features. Significant correlations with the expert-coded session level empathy scores are obtained. Therefore we suggest that for MI, therapist empathy can be partially evaluated by means of computational language modeling. Experimental results also imply that coders tend to assess therapist's empathy by accounting for salient empathic events in a session.

In Section II the two datasets and the details of observational behavior coding are introduced. In Section III we explain the way of building language models and feature extraction. Experiment results are reported in Section IV followed by discussion in Section V. Finally, we conclude the study in Section VI.

II. Data Sets

Both sets of data employed in the current study are from clinical trial studies using MI on substance use (drug abuse, alcohol use disorders, etc.) by college students. All sessions were manually transcribed, and only the therapist parts of transcripts are utilized. Similar text pre-processing are applied to the two sets of data: speaking turns are split into utterances either by the coder's segmentation in the first set, or by period in the transcripts in the second set; word-external punctuations, quotes, words within parentheses, and other special symbols are then removed; capitalized characters are converted into lowercase; hyphens, apostrophes, underscores, as well as special notes in brackets such as [laughs] are retained.

The first set comes from a part of three MI studies, referred to as ESP21, ESPSB and HMCBI. Three well-trained coders

evaluated these sessions based on audio and the original transcripts, following the Motivational Interviewing Skill Code (MISC) coding manual [7], which describes therapist and client behaviors at the utterance level, and assesses the therapist's overall competence. In addition, the coding team invented the code called "Brownie point", which was marked to an utterance whenever it locally exemplifies a type of global assessments of the therapist. "Empathy" is one of such assessments, with the coding instruction describing it as "therapists show active interest in making sure they understand what the client is saying". Brownie points make it easier to pinpoint typical empathic language perceived by the coders. In total, 28 sessions were analyzed. In order to maximize available training examples of empathic language, we collect empathy-coded utterances if any one of the coders put a marker of empathy on that utterance. Consequently an utterance is considered as non-empathic if none of the coders marked it as empathic. In total 854 empathic and 6439 non-empathic utterances are identified. We call this set as the MISC dataset.

The second set comes from a part of three other MI studies, referred to as ARC, iCHAMP and GOALS. The session level therapist coding scheme — Motivational Interviewing Treatment Integrity (MITI) [8] — was used to give global score of therapist empathy in Likert scale, ranging from 1 to 7 with 7 being highly empathic. Trained human coders used audio and original transcripts to perform the coding. In total, 88 sessions are collected, where the number of sessions having a score of 3 to 7 are 2, 24, 32, 29 and 1, respectively. Scores of 1 and 2 are not observed in the data. Therefore the empathy scores in this part are mainly 4, 5 and 6. We call this set as the MITI dataset.

In summary, the two datasets are presented in Table I.

TABLE I
Summary of MISC and MITI datasets

Dataset	Unit	Ratings
MISC	Utterance	Empathic: 854
		Non-empathic: 6439
MITI	Session	Empathy rating on a 1-7 Likert scale
		97% in the 4-6 range

III. Language Modeling

A. Maximum likelihood classifier

In a Maximum Likelihood sense, we build a classifier based on language model for the empathic and non-empathic classes. Let E and N denote the two classes above. Let an utterance formed by a word sequence $\{w_i|i = 1, 2, \cdots, l\}$ be denoted $\mathbf{w}$; the likelihoods based on N-gram language model of the two classes are $P(\mathbf{w}|E)$ and $P(\mathbf{w}|N)$. The decision of the classifier is as (1).

$$\mathbf{w} \in \arg\max_C P(\mathbf{w}|C), \quad C \in \{E, N\} \tag{1}$$

However, such a classifier would suffer from over-training because of small data size and disjoint samples in each class. To tackle this issue we utilize a language model trained on a large separate data set, and mix it with both classes in two steps. First, let the model derived from the large data set be denoted L, and a mixed *universal background model* (UBM) be B, the UBM is obtained by (2),

$$P(\mathbf{w}|B) = \lambda_1 P(\mathbf{w}|L) + \sum_{C=E,N} \frac{1-\lambda_1}{2} P(\mathbf{w}|C) \tag{2}$$

where λ_1 is the weight on L. Second, the UBM is mixed with both classes to obtain the final model as (3),

$$\tilde{P}(\mathbf{w}|C) = \lambda_2 P(\mathbf{w}|C) + (1-\lambda_2) P(\mathbf{w}|B), \; C \in \{E, N\} \tag{3}$$

where λ_2 is the weight on class E or N, and $\tilde{P}(\mathbf{w}|E)$, $\tilde{P}(\mathbf{w}|N)$ are the mixed language models for the two classes, respectively. The classifier is updated as in (4).

$$\mathbf{w} \in \arg\max_C \tilde{P}(\mathbf{w}|C), \quad C \in \{E, N\} \tag{4}$$

B. Session level feature extraction

The models generated above can also be used to extract session level lexical features of the therapist's overall empathy level. Let the set of such features be denoted $\mathbb{F}$. We define $d(\mathbf{w})$ to be the difference in log probability of an utterance given the two models as in (5).

$$d(\mathbf{w}) = \log \tilde{P}(\mathbf{w}|E) - \log \tilde{P}(\mathbf{w}|N) \tag{5}$$

We also use variants of (4) above to assign beliefs of empathy at the utterance level. For each of these variants, as in (6), $\mathbb{E}$ will denote the set of utterances from the whole session $\mathbb{U}$, that will be estimated as belonging to the empathy class.

The first feature $f_1 \in \mathbb{F}$ being considered is the sum of $d(\mathbf{w})$ for $\mathbf{w} \in \mathbb{U}$, interpreted as cumulative evidence of empathy in a session. Secondly, $d(\mathbf{w})$ can be binarized by its polarity, and summed to give $f_2 \in \mathbb{F}$, so that the session level feature is the count of decisions made at utterance level, with a threshold of 0 (can be viewed as decision with an equal prior). Thirdly, from a saliency point of view, we would like to accept $\mathbf{w} \in \mathbb{E}$ if $d(\mathbf{w})$ is larger than 0 by a moderate margin being δ_3. We denote this feature as $f_3 \in \mathbb{F}$. In addition, we take the ratio of empathic utterances in a session as f_4, that is equal to f_3 normalized by the number of utterances $|\mathbb{U}|$. We also design a feature f_5 that has varying $\delta_5(i) = \delta_5 \times l_i$, where l_i is the number of words of the i-th utterance of the session (utterance end symbol </s> included), so that longer utterances have higher threshold. Finally, a feature f_6 brings f_4 and f_5 together. The features f_1 to f_6 are summarized in (6).

$$\begin{aligned}
f_1 &= \sum_{\mathbf{w}\in\mathbb{U}} d(\mathbf{w}) \\
f_2 &= |\{\mathbf{w}|\mathbf{w} \in \mathbb{U},\ d(\mathbf{w}) > 0\}| \\
f_3 &= |\{\mathbf{w}|\mathbf{w} \in \mathbb{U},\ d(\mathbf{w}) > \delta_3\}| \\
f_4 &= |\{\mathbf{w}|\mathbf{w} \in \mathbb{U},\ d(\mathbf{w}) > \delta_4\}| \times \frac{1}{|\mathbb{U}|} \\
f_5 &= |\{\mathbf{w}|\mathbf{w} \in \mathbb{U},\ d(\mathbf{w}) > \delta_5 \times l_i\}| \\
f_6 &= |\{\mathbf{w}|\mathbf{w} \in \mathbb{U},\ d(\mathbf{w}) > \delta_6 \times l_i\}| \times \frac{1}{|\mathbb{U}|}
\end{aligned} \tag{6}$$

For simplicity, unless explicitly stated, we use δ_ϕ to denote any of δ_3 to δ_6. The value of δ_ϕ can be optimized on a development set. Note that reasonable values of δ_ϕ range from 0 to the maximum difference of log probability over the set of possible utterances. Through these bounds on δ_ϕ we can search for δ^*_ϕ that optimizes the effectiveness of the feature.

To evaluate, let $F = \{f_\phi(i)\}$ denote the feature stream of f_ϕ and $Y = \{y(i)\}$ denote session level empathy scores for K sessions ($i = 1, 2, \ldots, K$). In our study, we set the target function of the optimization to be in (7), where $\text{Corr}(F, Y)$ is the correlation between F and Y. The optimization is applied for all $\phi = 3 \ldots 6$.

$$\delta^*_\phi = \arg\max_{\delta_\phi} \text{Corr}(F, Y) \tag{7}$$

IV. Experiments

A. Empathic utterance classification — MISC dataset

We use the SRILM tool [9] to implement N-gram language model. The original model $P(\mathbf{w}|E)$ and $P(\mathbf{w}|N)$ are smoothed using Kneser-Ney algorithm. The switchboard text corpus [10] is used as the large dataset (L) towards generating the UBM (B).

On the MISC dataset, a 5-fold cross-validation is carried out, where the empathic and non-empathic utterances are equally split into 5 parts respectively, and in each fold one part of each class is held out. The remaining data are used to train the classifier as described in Section III-A.

For evaluation we will employ precision and recall in (8), where the C_E denotes the set of utterances marked by experts as empathic.

$$\text{precision} = \frac{|\{\mathbf{w}|\mathbf{w} \in \mathbb{E} \text{ and } \mathbf{w} \in C_E\}|}{|\mathbb{E}|} \qquad (8)$$
$$\text{recall} = \frac{|\{\mathbf{w}|\mathbf{w} \in \mathbb{E} \text{ and } \mathbf{w} \in C_E\}|}{|C_E|}$$

To test the effect of mixing parameters, we choose λ_1 and λ_2 from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, respectively. The empathy classification results on the held out sets using unigram, bigram or trigram features are shown in Figure 1, where points with the same λ_1 or λ_2 value are linked with a solid or dotted line, respectively.

We can observe that unigram features result in higher recall and lower precision, while bigram features are higher on precision but lower on recall. The performance using trigram features is worse than bigram. The highest F1-score of 0.56, with 0.48 precision and 0.66 recall, is achieved with $\lambda_1 = 0.5$, $\lambda_2 = 0.7$ and using bigram features.

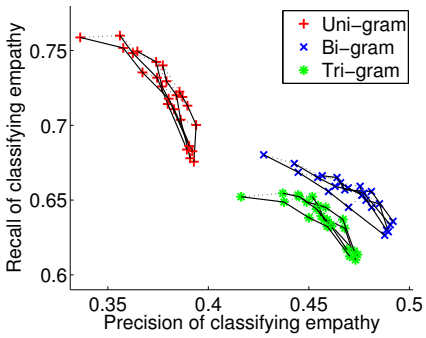

Fig. 1. *Precision and Recall of classifying empathic utterances with various λ_1 and λ_2*

The experiment shows that words in isolation, i.e. unigrams, do not separate empathic utterances as reliably as word usage in a context, i.e. bigrams. However we also observe that, likely due to increased sparsity issues resulting from their higher context representation, bigram and trigram features are not as robust in recall as unigram features and in addition trigram features perform worse in both precision and recall to bigram features.

B. Session level empathy — MITI dataset

In this experiment we test the effectiveness of the features proposed in Section III-B. Restricted by the size of MITI dataset, we use as much data as possible to optimize the δ_ϕ parameters through leave-one-out cross-validation (88 times of 87-dev, 1-test[1]). Using the test set we evaluate the correlation between Y and the features $\mathbb{F}$.

We take the bigram model learned on the whole MISC dataset with $\lambda_1 = 0.5$ and $\lambda_2 = 0.7$, i.e. the model yielding highest F1-score, as an example. To optimize δ_ϕ, we did a simple stepwise search with step size being 0.01 in each round (f_1 and f_2 do not require optimization). The baseline correlations of $f_1, \cdots, f_6$ and Y on all the sessions are obtained in Table II. f_1 fails to correlate significantly with Y; f_2 has a positive correlation at p-value = 0.001 significance. f_3 to f_6 are giving better correlation above 0.4. In figure 2 we plot the f_3 feature value on horizontal axis and the corresponding Y on vertical axis. Also we plot the histogram of f_3 value for different Y values. We can see there is a tendency of larger f_3 associated with larger session level empathy score.

[1]Note that training is already assumed through the MISC dataset models

TABLE II
Correlations of lexical features $\mathbb{F}$ and Y

Feature	f_1	f_2	f_3
Correlation	-0.11	0.35	0.41
p-value	0.3	1×10^{-3}	8×10^{-5}
Feature	f_4	f_5	f_6
Correlation	0.43	0.40	0.43
p-value	3×10^{-5}	1×10^{-4}	2×10^{-5}

Feature	$f_{1\ldots6}$	$f_{3\ldots6}$
Correlation	0.56	0.50
p-value	2×10^{-8}	1×10^{-6}

In addition, we are interested in the combined performance of using f_1 to f_6. Fitting the above $\mathbb{F}$ and Y to a linear regression model, the predicted $\hat{Y}$ has a correlation of 0.56 with Y. Comparing with the nested models only using f_3 to f_6 individually, the extended multi-variant model significantly improves accuracy under F-test at $\alpha = 0.05$. We should also note that the features in $\mathbb{F}$ are often highly correlated as they are not independently generated. For instance f_2 and f_5 have a correlation of 0.97. Therefore we adopted a Bayesian linear regression approach to mitigate the multi-collinearity issue, and achieved a correlation of 0.53 between $\hat{Y}$ and Y.

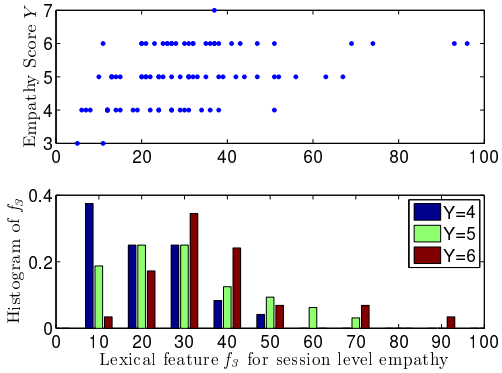

Fig. 2. *Feature f_3 and session level empathy score Y*

V. Discussion

A. High empathy words in unigram

To understand better the major distinguishing features of empathic and non-empathic language we report in this section the most discriminating words between the two models, ranked by the product of $d(\mathbf{w})$ and the number of word occurrence in MISC dataset, as denoted $D(\mathbf{w})$ in (9). With $\lambda_1 = 0.5$ and $\lambda_2 = 0.7$, the result is listed in Table III where words ranked by positive (empathic) and negative (non-empathic) $D(\mathbf{w})$ are displayed separately.

$$D(\mathbf{w}) = d(\mathbf{w}) \times number\_of\_occurrence(\mathbf{w}) \qquad (9)$$

We can see more second person pronouns and more reflective listening related words such as “sounds” for empathy; while on the other hand there are more first and third person pronouns, and more following-neutral words like “mm-hmm”. This matches highly

TABLE III
Words having prominent discriminative power

Empathy	Non-Empathy
you're, you, it, like, sounds, so, and, you've, your, of, that, to, it's, a, with, kind, not, really, for, kinda, time, friends, maybe	they, mm-hmm, what, we, alcohol, this, yeah, think, about, okay, drinks, right, if, do, is, that's, they're, b_a_c, us, umhum

reflective therapy-talk such as "It sounds like you're ...", with reflections being accepted in therapy as highly empathic language techniques.

B. Empathy perception as salient events

In Section IV-B the f_1 feature does not yield significant result, while features obtained via thresholding like f_3 is significantly correlated with overall empathy. One interpretation is that the degree of empathy in a session perceived by the coder is not precisely the cumulative level of empathy of each utterance, but enough occurrences of salient empathic utterances act as "highlights" to strengthen the coder's decision. This matches existing theories of perception such as the Gestalt Principle Theory of Perception [11]. In addition it's worth noting that our non-empathic training samples are more of generic and neutral language rather than the exact opposite of empathy, so higher probability on the non-empathy model does not mean highly against empathy.

C. Related work on modeling empathy

There have been a few studies of computational models of empathy in the literature. In [12] the authors constructed a system of virtual environment involving the user and a virtual agent. In training mode, a human trainer guided the virtual agent to act in an empathic manner. In test mode, the system decides when and how should the virtual agent act in an empathic manner. Timing, location and intention information were employed as features within the virtual environment. Naive Bayes and Decision Tree models were adopted in learning. Experiments showed the system could provide the basis of empathic behavior control of the virtual agent. In [13] the authors suggested that the occurrence and attribute of emotional interaction (i.e., empathic, antipathetic or unconcerned) are related to facial expression and gaze in multi-person interaction. Computer vision techniques were used to detect "who is facing whom and when", and the empathy level notes were provided by human evaluators. The authors built a Bayesian learning model to estimate level of empathy via the extracted cues. Experiments showed that the system was able to infer the empathy behavior.

D. Future work: towards an evaluator model of empathy

We have analyzed language modeling of empathy in a Maximum Likelihood sense. In fact, there are many more aspects that can be incorporated. For example, the client's language is not utilized in this study. A more complete model should consider the therapist's empathic language in the context of the conversation with the client. As studied in [3], there are opportunities of expressing empathic language for the therapist within the context. By tracking or hypothesizing such opportunities, one could get a more accurate measure of how well the therapist is doing. As shown in the above discussion of word use, empathic language is often related to reflection to the client's talk. Locating reflections [14] by the therapist might be helpful for evaluating empathy.

Moreover, recall that empathy is not only expressed in language, but also via many other modalities, such as the way of saying as acoustic features, the body gesture and motion, facial expression and eye contact. A better evaluator model of empathy should ideally incorporate these modalities and conduct reasoning in the context of the conversation. For example, we have successfully used such BSP approaches in behavioral coding of distressed couple interactions [5].

VI. Conclusion

In this paper we introduced empathy as an important aspect in social communication, especially in medical and psychotherapy applications. For psychotherapy based on Motivational Interview, the characteristics of empathic and non-empathic behavior are learned with N-gram language models. A language-based classifier of empathic versus non-empathic utterances is proposed in Maximum Likelihood sense. High recall is obtained with unigram while bigram achieved the highest F1-score. Based on the language model, a group of lexical features are proposed, and tested by the correlation with expert-coded session level empathy scores. Combined features achieved a correlation of 0.56 between predicted session level empathy scores and the expert-coded ones. The study suggests that in the scenario of psychotherapy where language use is constrained by the application, computational language modeling can provide useful insights into the expressed empathy behavior of therapists. Moreover, experiments show that human coders tend to assess session level empathy as a gestalt of salient empathic behavior.

References

[1] J. Decety, "A social cognitive neuroscience model of human empathy," *Social neuroscience: Integrating biological and psychological explanations of social behavior*, pp. 246–270, 2007.

[2] P. Bellet and M. Maloney, "The importance of empathy as an interviewing skill in medicine," *Journal of the American Medical Association*, vol. 266, no. 13, pp. 1831–1832, 1991.

[3] A. Suchman, K. Markakis, H. Beckman, and R. Frankel, "A model of empathic communication in the medical interview," *Journal of the American Medical Association*, vol. 277, no. 8, pp. 678–682, 1997.

[4] P. Georgiou, M. Black, A. Lammert, B. Baucom, and S. Narayanan, "'that's aggravating, very aggravating': Is it possible to classify behaviors in couple interactions using automatically derived lexical features?" in *Proc. ACII*. Springer, 2011, pp. 87–96.

[5] M. Black, A. Katsamanis, B. Baucom, C. Lee, A. Lammert, A. Christensen, P. Georgiou, and S. Narayanan, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Communication*, 2011.

[6] C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Proc. IS*, 2010.

[7] W. Miller, T. Moyers, D. Ernst, and P. Amrhein, "Manual for the motivational interviewing skill code (misc), version 2.1," *Substance Abuse and Addiction (CASAA), University of New Mexico.*, 2008.

[8] T. Moyers, T. Martin, J. Manuel, and W. Miller, "The motivational interviewing treatment integrity (miti) code: Version 2.0," *Unpublished. Albuquerque, NM: University of New Mexico, Center on Alcoholism, Substance Abuse and Addictions*, 2008.

[9] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Proc. ICSLP*, vol. 2, 2002, pp. 901–904.

[10] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP*, vol. 1. IEEE, 1992, pp. 517–520.

[11] G. Humhprey, "The psychology of the gestalt," *Journal of Educational Psychology*, vol. 15(7), pp. 401–412, 1924.

[12] S. McQuiggan and J. Lester, "Modeling and evaluating empathy in embodied companion agents," *International Journal of Human-Computer Studies*, vol. 65, no. 4, pp. 348–360, 2007.

[13] S. Kumano, K. Otsuka, D. Mikami, and J. Yamato, "Analyzing empathetic interactions based on the probabilistic modeling of the co-occurrence patterns of facial expressions in group meetings," in *Proc. FG*. IEEE, 2011, pp. 43–50.

[14] D. Can, P. Georgiou, D. Atkins, and S. Narayanan, "A case study: Detecting reflections with linguistic features in motivational interviewing based psychological therapy," in *Proc. IS*, 2012.