INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Modeling Interpersonal Linguistic Coordination in Conversations using Word Mover's Distance

Md Nasir[1], *Sandeep Nallan Chakravarthula*[1], *Brian RW Baucom*[2], *David C Atkins*[3], *Panayiotis Georgiou*[1], *Shrikanth Narayanan*[1]

[1]University of Southern California, Los Angeles, CA, USA
[2]University of Utah, Salt Lake City, UT, USA
[3]University of Washington, Seattle, WA, USA

{mdnasir,nallanch}@usc.edu, brian.baucom@utah.edu, datkins@uw.edu,
georgiou@sipi.usc.edu, shri@ee.usc.edu

Abstract

Linguistic coordination is a well-established phenomenon in spoken conversations and often associated with positive social behaviors and outcomes. While there have been many attempts to measure lexical coordination or entrainment in literature, only a few have explored coordination in syntactic or semantic space. In this work, we attempt to combine these different aspects of coordination into a single measure by leveraging distances in a neural word representation space. In particular, we adopt the recently proposed Word Mover's Distance with *word2vec* embeddings and extend it to measure the dissimilarity in language used in multiple consecutive speaker turns. To validate our approach, we apply this measure for two case studies in the clinical psychology domain. We find that our proposed measure is correlated with the therapist's empathy towards their patient in Motivational Interviewing and with affective behaviors in Couples Therapy. In both case studies, our proposed metric exhibits higher correlation than previously proposed measures. When applied to the couples with relationship improvement, we also notice a significant decrease in the proposed measure over the course of therapy, indicating higher linguistic coordination.

Index Terms: entrainment, Word Mover's Distance, linguistic coordination, empathy, outcome

1. Introduction

When people engage in conversations in social settings, they tend to coordinate with each other and show similar behavior in various modalities. This tendency, known as entrainment or coordination, is exhibited through facial expressions [1], head-motion [2], vocal patterns (vocal entrainment) [3, 4], as well as the use of language (linguistic coordination) [5]. Linguistic coordination is a well-established phenomenon in both spoken and written communication that has many collaborative benefits. It is often associated with a wide range of positive social behaviors and outcomes, such as task success in collaborative games [6, 7], building effective dialogues [8] and rapport [9], engagement in tutoring scenario [10], successful negotiation [11].

Understanding linguistic coordination and quantifying it is beneficial in characterization of interpersonal behavior in psychotherapy, and in monitoring the quality and efficacy of therapy [12, 13]. Another potential application lies in spoken dialog systems and conversational agents, where the system can learn to use linguistic coordination to communicate efficiently with the human user and create a common ground [7].

According to Pickering and Garrod's model [5], there exist several different components in linguistic coordination – lexical, syntactic and semantic. Among these lexical entrainment has been arguably the focus of the most attention, primarily in psycholinguistics [14, 15]. While it is a complex and multifaceted phenomenon, a number of studies have explored specific forms of lexical entrainment, such as linguistic style matching [16], similarity in choice of high frequency words [6], similarity in referring expressions [15], similarity in style words [17]. Researchers in computational linguistics also tried to quantitatively measure lexical entrainment in conversational settings. For example, [6] used a unigram model of different classes of words and measured lexical entrainment as the cumulative difference in unigram scores for the interlocutors.

However, the majority of the computational approaches for measuring linguistic coordination has been limited to lexical entrainment, agnostic to coordination in the semantic space or syntactic structures. Coordination in semantics is closely related to cohesion [18], another mechanism in linguistics which ties together different words used in continuation of a shared context. Approaches towards quantification of cohesion primarily have been used in tasks like text classification and discourse segmentation [19]. In these applications, however, cohesion is defined within a document, as opposed to the cohesion between the interlocutors in dyadic conversations which we are interested in. There have been only a few attempts to model the latter by exploring the relation between synonymous words (*e.g.,* via WordNet) used by different speakers in the domain of intelligent tutor systems [10, 20]. However, this body of work suffers from the limitation that two words might be semantically or syntactically related even without being synonyms. Further, using any of the lexical entrainment or cohesion measures alone does not provide a complete representation of linguistic coordination.

Addressing the aforementioned limitations and drawing inspiration from the recent success of neural word embeddings, we adopt a distance measure known as Word Mover's Distance (WMD) [21] and extend it to compute a distance that captures linguistic coordination. The primary novelty in our work is in jointly integrating multiple aspects of coordination into a single measure. In our framework, we also propose to measure the coordination locally and then normalize it globally to account for the individual tendency of coordination. We experimentally validate our measure in relation to the therapist's empathy towards their patient in Motivational Interviewing as well as outcome and affective behaviors in Couples Therapy.

http://dx.doi.org/10.21437/Interspeech.2019-1900

2. Lexical Similarity in Conversations

Word Mover's Distance, originally proposed as a lexical distance between two documents, is used as a building block to measure lexical distance between two interlocutors in a dyadic conversation. In this section, first we discuss the basics of Word Mover's Distance and then propose how the distance between utterances of two interlocutors could be used as conversational distance measure that can capture lexical and semantic dissimilarity.

2.1. Word Mover's Distance (WMD)

Word Mover's Distance (WMD) was introduced by Kusner *et al.* [21], as a distance measure between text documents. The measure is based on the concept of neural word embeddings, which provide distributed vector representations of words in a document. Although any neural word embedding could be used in measuring WMD, it was originally proposed using one of the most popular word embeddings, *word2vec* [22]. *word2vec* has been shown to contain semantic and syntactic information [22], making WMD suitable for capturing different aspects of linguistic coordination. Unlike the original WMD paper, we include stop words (which do not carry much semantic information) in our framework, in order to capture lexical entrainment patterns of using similar high-frequency and style words. WMD is essentially a bag-of-words approach where each document is a collection of words represented as vectors in the embedding space. In principle, it can be interpreted as the minimum transport cost to reach the embedded words in a document from the embedded words of another document. Inherently this measure relies on the individual distances of pairs of words in the vector space, as building blocks. For a pair of words, w_i and w_j, the Euclidean distance between their embedding vectors is computed as the first step, $\mathbf{v_i} = e(w_i)$ and $\mathbf{v_j} = e(w_j)$,

$$d(w_i, w_j) = \|\mathbf{v}_i - \mathbf{v}_j\| \tag{1}$$

Based on this, the distance between a pair of utterances U_1 and U_2 is formulated as follows:

$$\mathrm{WMD}(U_1, U_2) = \min_{\mathbf{T} \geq 0} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{T}_{ij} d(w_i, w_j) \tag{2}$$

$$\text{subject to} \quad \sum_{j=1}^{n} \mathbf{T}_{ij} = \frac{c_i^1}{n} \quad \forall i \in \{1, \ldots, m\},$$

$$\text{and} \quad \sum_{i=1}^{m} \mathbf{T}_{ij} = \frac{c_j^2}{m} \quad \forall j \in \{1, \ldots, n\},$$

where m and n are the number of unique words in U_1 and U_2 respectively and c_i^k is the frequency of w_i in U_k. The computation of WMD involves a constrained optimization problem of finding an optimal flow matrix $\mathbf{T}$ which can be solved using many exact and approximate techniques. In fact, this is a special case of *earth mover's distance* computation, a widely-known transportation problem [23]. In Figure 1, we illustrate how WMD between two utterances is computed in the vector space of word embeddings (only two dimensions are shown for interpretability). The optimal selection of $\mathbf{T}$ could be interpreted as finding ties between neighboring words in the vector space, as seen in the figure.

Although WMD was originally introduced for documents, more recently it has been also applied for sentences [24], and in this work, we use it for utterances.

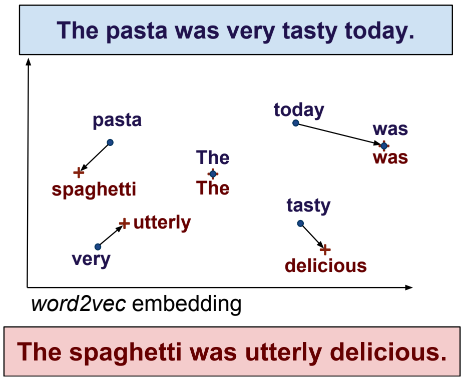

Figure 1: *Illustration of WMD (each word from one utterance is mapped to the most similar word in the other utterance)*

2.2. Conversational Linguistic Distances

As discussed earlier, WMD can provide a measure of linguistic difference between two utterances. Here we describe how it is extended to the distance measure capturing linguistic coordination, which we name *Conversational Linguistic Distance* (CLiD). More specifically, we propose an *unnormalized* and a *normalized* distance (uCLiD and nCLiD).

2.2.1. Local Interpersonal Distance

Although linguistic coordination occurs at multiple levels, we focus on capturing it at a local scale, *i.e.,* between consecutive turns of the interlocutors. The other alternative is to measure the coordination globally by considering all the words used by each of the interlocutors as a single document and computing the distance between them. While similar approaches have been adopted in prior works on lexical entrainment [6], the coarse resolution of such a measure can potentially fail to capture the dynamics of the conversation.

On the other hand, measuring the distance between one speaker turn and the immediate next one is a simple local measure which is appealing for our purpose. However, local coordination is not necessarily expressed in the immediate response to the primary speaker's turn; rather it might be sustained and exhibited after a few turns [25]. Hence, we propose a scheme where we consider a predefined number of turns (defined as context length) in response to the utterance of the primary speaker (referred to as anchor), and choose the minimum of the distances of every pair formed by the anchor utterance and a response. This can be interpreted as the maximum coordination that is exhibited towards the primary speaker by their interlocutor in the causal vicinity of the original utterance. In a similar approach, [26] considered a predefined time window (as opposed to fixed number of turns) as the context length to find instances of syntactic coordination.

Let us consider the scenario where two interlocutors A and B converse with each other and each of them takes N number of turns. $A_1, A_2, .., A_N$ and $B_1, B_2, .., B_N$ represent the utterances of A and B respectively. Given a context length k, for every anchor utterance A_i, we compute a distance $d_i^{A \to B}$ over next k number of utterances by B following A_i as follows:

$$d_i^{A \to B} = \min_{i \leq j \leq i+k-1 \leq N} \mathrm{WMD}(A_i, B_j) \tag{3}$$

It should be noted that we obtain two sequences of directional distance measures for the entire session, $\{d_i^{A \to B}\}$ and $\{d_i^{B \to A}\}$, due to the asymmetric nature of Equation (3).

2.2.2. *Session-level measures*

Although local distance measures provide a good characterization of the interpersonal coordination that happens throughout the course of conversation, an aggregated session-level measure obtained from the local distances could be more useful for session-level analysis in applications like behavioral analyses. We simply take an average of the local distances defined in Equation (3) over the whole session to compute the session-level measures, which we call *unnormalized Conversational Linguistic Distance* (uCLiD):

$$\text{uCLiD} = \frac{1}{N}\sum_{i=1}^{N} d_i^{A\rightarrow B} \tag{4}$$

In this equation, only uCLiD for $A \rightarrow B$ has been shown, that captures interlocutor A's coordination with B; similarly $B \rightarrow A$ can be computed. While the uCLiD measure provides how much overall linguistic coordination occurs between interlocutors in a conversation, it is also influenced by the nature of the conversation – whether it is a structured conversation on pre-decided topic or an unrestricted spontaneous interaction, or something in between. It can be also affected by the extent to which the interlocutors tend to use similar language in a conversation as a whole, as a result of coordinating to their own language. To account for these phenomena, we use a normalized distance which attempts to provide a more suitable measure for applications where the nature of the conversation is not important. We draw inspiration from a similar approach by Jones *et al.* [27], where they compute a factor called *Zelig Quotient* for normalization. In our work, we first define a normalization factor α, computed as the average pairwise WMD measure throughout the session, including within and across interlocutors. Next, the normalized distance measure, which we term as *normalized Conversational Linguistic Distance* (nCLiD) is computed by dividing uCLiD by α, as follows:

$$\text{nCLiD} = \frac{\text{uCLiD}}{\alpha}, \tag{5}$$

$$\begin{aligned}\text{where } \alpha = {} & \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\text{WMD}(A_i, A_j) \\ & + \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\text{WMD}(B_i, B_j) \\ & + \frac{2}{N(N+1)}\sum_{i=1}^{N}\sum_{j=i}^{N}\text{WMD}(A_i, B_j)\end{aligned} \tag{6}$$

In the RHS of Equation (6), the first two terms are the average WMD within A and within B, which are related to the tendency to change their language throughout the conversation. The third term represents the overall tendency of each interlocutor to accommodate the other.

3. Datasets

Two datasets are used in this work: a corpus consisting of five independent clinical studies in addiction counseling (Motivational Interviewing corpus) and another corpus consisting of interactions of married couples undergoing marital therapy (Couples Therapy corpus).

3.1. Motivational Interviewing corpus

This corpus consists of therapist-patient interactions in Motivational Interviewing (MI), a form of addiction counseling in psychotherapy. In each interview, the aim of the therapist is to help the patient, who is seeking therapy for substance addiction, make behavioral changes by resolving ambivalence about their problems. There are 145 interactions, in total, collected from the five clinical studies: ARC, ESPSB, ESB21, iCHAMP, HMCBI [28]. The interactions, which range from 20 minutes to an hour, take place between therapists and real patients struggling with alcohol, marijuana and poly-drug addiction.

Each interaction was recorded on tape and manually transcribed and annotated for speaker labels, turn timings, back-channels, disfluencies, etc. In addition, each therapist was assigned an overall, session-level rating for the behavior code *empathy* based on the Motivational Interviewing Treatment Integrity (MITI) [29] manual. The rating was performed on a Likert Scale from 1 to 7, where low (high) values indicated low (high) levels of empathy exhibited by the therapist.

3.2. Couples Therapy corpus

The second dataset used in this work was collected as part of longitudinal study conducted by University of California, Los Angeles and University of Washington [30]. 134 seriously and chronically distressed heterosexual couples received therapy and participated in sessions where each spouse discussed with their partner one problem relevant to their relationship, without any therapist or research staff present. There is a total of 574 such sessions, recorded at three different points of time over a span of two years while undergoing therapy (before therapy, after 26 weeks and 2 years since the beginning of the therapy). Along with audio-visual recordings, the corpus also includes manual transcripts with speaker labels of the conversations.

For each session, both of the spouses are evaluated with 32 session-level behavioral codes using two separate coding schemes. 19 of the codes are based on the Social Support Interaction Rating System (SSIRS) while 13 of them follow The Couples Interaction Rating System (CIRS). All of these codes are rated by three to four trained annotators for each session on a scale from 1 to 9. In this work, our focus lies on analyzing only two codes from the SSIRS system – *Global Positive Affect* and *Global Negative Affect*. Finally, the corpus also includes the therapy outcomes of the couples as a measure of their relationship quality relative to the beginning of the therapy. Rated on two occasions (26 weeks and/or 2 years), which we refer to as *post-therapy* sessions, the outcome is rated on a 4-point scale; 1 (deterioration), 2 (no change), 3 (partial recovery), and 4 (complete recovery).

4. Experiments

We applied the proposed measure in the two case studies using the datasets described in Section 3. In this section, we describe the correlation analysis experiments conducted to indirectly validate our proposed measures.

4.1. Baselines

We use a number of baseline methods to compare with the proposed method:

- Turn-level lexical similarity based on TF-IDF [31],
- Cohesion (distance) measure based on WordNet [10],

- Global WMD measured between the language of the interlocutors taken together, as described in Section 2.2.1

4.2. Case Study 1: Empathy in Motivational Interviews

Deemed an important interpersonal behavior in counseling-based psychotherapy, empathy has been shown to be positively associated with entrainment both in domain theory [32] and computational studies [33, 34]. In this case study, we compute Spearman's ρ correlation between the proposed linguistic coordination measures (uCLiD and nCLiD) and empathy ratings. Due to the asymmetric nature of the proposed measure, we obtain each of these measures in two directions–the *patient-to-therapist* and *patient-to-therapist*. Since empathy is a behavior expressed by therapist, intuitively it should not be affected by how much coordination the patients exhibits. As a verification, we found no significant correlation between the therapist-to-patient distance (using nCLiD measure) and empathy ($\rho = 0.0521, p = 0.4344$). Hence we consider only patient-to-therapist coordination distance, focusing only on the coordination exhibited by the therapist. We empirically set the context length parameter of our measure as $k = 6$ and use a 300-dimensional pre-trained model for *word2vec* (trained on 3 million words from Google News). We also report the p-values against the null hypothesis H_0 that there is no monotonic (rank-ordered) association between empathy and the candidate measure. We repeat the same procedure for the baselines as well.

Table 1: *Correlation between empathy and various coordination measures*

Measure	Spearman's correlation	
	ρ	p-value*
uCLiD	-0.2283	0.0103
nCLiD	$\mathbf{-0.2639}$	0.0026
†TF-IDF [31]	0.1152	0.1675
WordNet [10]	-0.0952	0.2546
global WMD	-0.1710	0.0398

From the results shown in Table 1, we can observe that both the normalized and the unnormalized measure (uCLiD and nCLiD, respectively) exhibit stronger correlation than the baselines. We also notice the improvement from normalization as nCLiD turns out to be the most highly correlated measure. The negative sign of the correlation values is justified for the proposed measures since we expect sessions with higher empathy to have higher coordination, and hence, lower distance. We also observe p-values lower than 0.05 indicating statistically significant association between empathy and the proposed measures.

4.3. Case Study 2: Couples Therapy

4.3.1. Individual behavioral codes

In the Couples Therapy domain, we first explore the possible association of linguistic coordination with *positive* and *negative* affect. We adopt the same context length parameter value for our measures ($k = 6$) and use the same baselines for comparison as used in the previous case study. We consider the coordination exhibited by a subject (husband or wife) with their spouse for the behavior ratings of the former. For example, as far as the husband's *positive* affective behavior is concerned, we only analyze how much the husband coordinated with respect to the wife during the session.

Table 2: Correlation between various coordination measures and affective behaviors (*positive* and *negative)*

Measure	*positive*		*negative*	
	ρ	p-value*	ρ	p-value*
uCLiD	-0.2903	9.9×10^{-5}	0.3142	3.4×10^{-8}
nCLiD	$\mathbf{-0.3068}$	1.2×10^{-7}	$\mathbf{0.3371}$	2.1×10^{-10}
†TF-IDF [31]	0.1542	0.0001	-0.2119	2×10^{-4}
WordNet [10]	-0.0847	0.0020	0.0952	0.0005
global WMD	-0.1310	0.0001	0.1556	0.0001

The results in Table 2 show that we obtain higher correlation values for our proposed measures than the baselines and that the normalized measure again exhibited the strongest correlation. Judging by the sign of ρ, coordination distance is higher for subjects with lower *positive* affect and lower for subjects with lower *negative* affect, which is consistent with literature associating entrainment with behavior [4].

4.3.2. Therapy outcome

We hypothesize that the coordination distance between both spouses (measured by the average of husband-to-wife and wife-to-husband distances) decreases in the post-therapy session with respect to the pre-therapy if they had fully recovered (outcome rating "4"). We conduct a paired Wilcoxon signed rank test against the null hypothesis H_0 that both pre- and post-therapy measures come from the same distribution. We obtain $p = 0.0125$ for uCLiD and $p = 0.0181$ for the nCLiD measure. This indicates a statistically significant ($p < 0.05$) observation that the couples who had recovered also exhibited lower coordination distance, or in other words, higher linguistic coordination after therapy.

5. Conclusion and Future Work

In this work, we present a novel distance measure to quantify linguistic coordination in dyadic conversations. Equipped with neural word embeddings, our proposed measure can potentially capture different aspects of linguistic coordination (lexical, semantic and syntactic). From the experiments performed in the two case studies, we establish the usefulness of the measure in capturing interpersonal behavioral information. In the future, we intend to study the effect of the context length parameter on our measure. We could use more recent and potentially more powerful neural word embedding techniques (such as BERT, ELMo) instead of *word2vec* in a similar framework as presented in this paper. Motivated by the efficacy of the neural word embeddings in relation to linguistic coordination, we would also like to explore models to jointly learn the embedding that encodes shared linguistic information between the interlocutors, similar to [35]. We would also like to investigate linguistic coordination *in-the-wild* through ASR transcripts using embeddings such as *conf2vec* [36]. Another possible research direction is to investigate modeling a fused measure combining linguistic and vocal coordination.

6. Acknowledgements

This research was supported in part by NSF, NIH, and DOD.

*$p < 0.05$ indicates statistically significant correlation

† similarity measure while other measures are distances

7. References

[1] F. Ramseyer and W. Tschacher, "Nonverbal synchrony of head- and body-movement in psychotherapy: different signals have different associations with outcome," *Frontiers in psychology*, vol. 5, p. 979, 2014.

[2] B. Xiao, P. G. Georgiou, C.-C. Lee, B. Baucom, and S. S. Narayanan, "Head motion synchrony and its correlation to affectivity in dyadic interactions," in *2013 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2013, pp. 1–6.

[3] M. Nasir, B. Baucom, S. Narayanan, and P. Georgiou, "Towards an unsupervised entrainment distance in conversational speech using deep neural networks," in *Interspeech / arXiv:1804.08782*, 2018.

[4] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions," *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, 2014.

[5] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[6] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers*. Association for Computational Linguistics, 2008, pp. 169–172.

[7] J. Lopes, M. Eskenazi, and I. Trancoso, "From rule-based to data-driven lexical entrainment models in spoken dialog systems," *Computer Speech & Language*, vol. 31, no. 1, pp. 87–112, 2015.

[8] R. Porzel, A. Scheffler, and R. Malaka, "How entrainment increases dialogical effectiveness," in *Proceedings of the IUI*, vol. 6. Citeseer, 2006, pp. 35–42.

[9] J. Cassell, A. J. Gill, and P. A. Tepper, "Coordination in conversation and rapport," in *Proceedings of the workshop on Embodied Language Processing*. Association for Computational Linguistics, 2007, pp. 41–50.

[10] A. Ward and D. Litman, "Measuring convergence and priming in tutorial dialog," *University of Pittsburgh, Tech. Report*, 2007.

[11] P. J. Taylor and S. Thomas, "Linguistic style matching and negotiation outcome," *Negotiation and Conflict Management Research*, vol. 1, no. 3, pp. 263–281, 2008.

[12] S. Narayanan and P. G. Georgiou, "Behavioral Signal Processing: deriving human behavioral informatics from speech and language," *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, vol. 101, no. 5, p. 1203, 2013.

[13] S. L. Koole and W. Tschacher, "Synchrony in psychotherapy: A review and an integrative framework for the therapeutic alliance," *Frontiers in psychology*, vol. 7, p. 862, 2016.

[14] S. Garrod and A. Anderson, "Saying what you mean in dialogue: A study in conceptual and semantic co-ordination," *Cognition*, vol. 27, no. 2, pp. 181–218, 1987.

[15] S. E. Brennan, "Lexical entrainment in spontaneous dialog," *Proceedings of ISSD*, vol. 96, pp. 41–44, 1996.

[16] K. G. Niederhoffer and J. W. Pennebaker, "Linguistic style matching in social interaction," *Journal of Language and Social Psychology*, vol. 21, no. 4, pp. 337–360, 2002.

[17] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais, "Mark my words!: linguistic style accommodation in social media," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 745–754.

[18] M. A. K. Halliday and R. Hasan, *Cohesion in english*. Routledge, 2014.

[19] O. Manabu and H. Takeo, "Word sense disambiguation and text segmentation based on lexical cohesion," in *Proceedings of the 15th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 1994, pp. 755–761.

[20] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-metrix: Analysis of text on cohesion and language," *Behavior research methods, instruments, & computers*, vol. 36, no. 2, pp. 193–202, 2004.

[21] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International Conference on Machine Learning*, 2015, pp. 957–966.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[23] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Sixth International Conference on Computer Vision*. IEEE, 1998, pp. 59–66.

[24] F. Ren and N. Liu, "Emotion computing using word mover's distance features based on ren_cecps," *PloS one*, vol. 13, no. 4, p. e0194136, 2018.

[25] M. J. Pickering and S. Garrod, "Alignment as the basis for successful communication," *Research on Language & Computation*, vol. 4, no. 2, pp. 203–228, 2006.

[26] D. Reitter, F. Keller, and J. D. Moore, "Computational modelling of structural priming in dialogue," in *Proceedings of the Human Language Technology Conference of the NAACL*. Association for Computational Linguistics, 2006, pp. 121–124.

[27] S. Jones, R. Cotterill, N. Dewdney, K. Muir, and A. Joinson, "Finding zelig in text: A measure for normalising linguistic accommodation," in *Proceedings of COLING 2014*, 2014, pp. 455–465.

[28] D. C. Atkins, M. Steyvers, Z. E. Imel, and P. Smyth, "Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification," *Implementation Science*, vol. 9, no. 1, p. 49, 2014.

[29] T. B. Moyers, T. Martin, J. K. Manuel, and W. R. Miller, "The motivational interviewing treatment integrity (MITI) code: Version 2.0."

[30] A. Christensen, D. C. Atkins, S. Berns, J. Wheeler, D. H. Baucom, and L. E. Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples." *Journal of consulting and clinical psychology*, vol. 72, no. 2, p. 176, 2004.

[31] N. Liebman and D. Gergle, "Capturing turn-by-turn lexical similarity in text-based communication," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 2016, pp. 553–559.

[32] S. D. Preston and F. B. De Waal, "Empathy: Its ultimate and proximate bases," *Behavioral and brain sciences*, vol. 25, no. 1, pp. 1–20, 2002.

[33] B. Xiao, P. G. Georgiou, Z. E. Imel, D. C. Atkins, and S. Narayanan, "Modeling therapist empathy and vocal entrainment in drug addiction counseling." in *INTERSPEECH*, 2013, pp. 2861–2865.

[34] S. P. Lord, E. Sheng, Z. E. Imel, J. Baer, and D. C. Atkins, "More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client," *Behavior therapy*, vol. 46, no. 3, pp. 296–303, 2015.

[35] S.-Y. Tseng, B. R. Baucom, and P. G. Georgiou, "Approaching human performance in behavior estimation in couples therapy using deep sentence embeddings." in *INTERSPEECH*, 2017, pp. 3291–3295.

[36] P. G. Shivakumar and P. Georgiou, "Confusion2vec: Towards enriching vector space word representations with representational ambiguities," *arXiv preprint arXiv:1811.03199*, 2018.