LINGUISTIC ENTRAINMENT IN MULTI-PARTY SPOKEN DIALOGUES

by

Zahra Rahimi

B.Sc. in Computer Engineering -Software, University of Tehran, 2009

M.Sc. in Software Engineering, University of Tehran, 2012

M.Sc. in Intelligent Systems Program, University of Pittsburgh, 2015

Submitted to the Graduate Faculty of
the School of Computing and Information Sciences, Intelligent Systems Program in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION SCIENCES

This dissertation was presented

by

Zahra Rahimi

It was defended on

May 28th 2019

and approved by

Diane Litman, Intelligent Systems Program

Kevin Ashley, Intelligent Systems Program

Rebecca Hwa, Intelligent Systems Program

Louis-Philippe Morency, School of Computer Science, Carnegie Mellon University

Dissertation Director: Diane Litman, Intelligent Systems Program

LINGUISTIC ENTRAINMENT IN MULTI-PARTY SPOKEN DIALOGUES

Zahra Rahimi, PhD

University of Pittsburgh, 2019

Entrainment is the propensity of speakers to begin behaving like one another in conversations. Evidence of entrainment has been found in multiple aspects of speech, including acoustic-prosodic and lexical. More interestingly, the strength of entrainment has been shown to be associated with numerous conversational qualities, such as social variables. These two characteristics make entrainment an interesting research area for multiple disciplines, such as natural language processing and psychology. To date, mainly simple methods such as unweighted averaging have been used to move from pairs to groups, and the focus of prior multi-party work has been on text rather than speech (e.g., Wikipedia, Twitter, online forums, and corporate emails). The focus of this research, unlike previous studies, is multi-party spoken dialogues. The goal of this work is to develop, validate, and evaluate multi-party entrainment measures that incorporate characteristics of multi-party interactions, and are associated with measures of team outcomes.

In this thesis, first, I explore the relation between entrainment on acoustic-prosodic and lexical features and show that they correlate. In addition, I show that a multi-modal model using entrainment features from both of these modalities outperforms the uni-modal model at predicting team outcomes. Moreover, I present enhanced multi-party entrainment measures which utilize dynamics of entrainment in groups for both global and local settings. As for the global entrainment, I present a weighted convergence based on group dynamics. As the first step toward the development of local multi-party measures, I investigate whether local entrainment occurs within a time-lag in groups using a temporal window approach. Next, I propose a novel approach to learn a vector representation of multi-party local entrainment by

encoding the structure of the presented multi-party entrainment graphs. The positive results of both the global and local settings indicate the importance of incorporating entrainment dynamics in groups. Finally, I propose a novel approach to incorporate a team-level factor of gender-composition to enhance multi-party entrainment measures. All of the proposed works are in the direction of enhancing multi-party entrainment measures with the focus on spoken dialogues although they can also be employed on text-based communications.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to specially thank my advisor, Dr. Diane Litman. The lessons I learned from you on how to do research, write a paper, and also about time management helped me through this journey. I would like to extend my sincere gratitude to the members of my dissertation committee, Dr. Louis-Philippe Morency, Dr. Rebecca Hwa, and Dr. Kevin Ashley, for their invaluable advice. I am also grateful to Susannah Paletz from whom I learned a lot while working on this dissertation. I thank Azadeh Shakery, my advisor at University of Tehran, under her guidance I first started doing research in information retrieval and natural language processing.

My deepest gratitude goes to my mother, Shamsi. Without her unconditional love and constant support, I would have never achieved this dream. Words cannot express my thankfulness. I would like to thank my father, Mohammad. I would like to thank my amazing older sister and brother, Sara and Morteza, for always being supportive. I thank my little sister, Maryam, who has always been my best friend.

Many special thanks to my wonderful friends, Huma Hashemi and Mehdi Pakdaman. I share so many good memories with them, including co-authoring my first paper at University of Pittsburgh, attending conferences together, and many lunchtime discussions. I am deeply grateful to my amazing friend, Majid Darvishan, who have been there for me in difficult times when I needed help the most. I am grateful to my friends at ISP, CS, and Pitt for making these years at Pittsburgh such a memorable experience. Thanks to Fattaneh Jabbari, Amin Tajgardoon, Omid Kashefi, Faezeh Movahedi, Jeya Balaji Balasubramanian, Jaromir Savelka, Jana Savelkova, Gaurav Trivedi, Fan Zhang, Wencan Luo, Huy Nguyen, Luca Lugini, Tazin Afrin, Mingzhi Yu, Haoron Zhang, Ahmed Magooda, and ChangSheng Liu.

1.0 INTRODUCTION

1.1 MOTIVATION

Entrainment[1] is the propensity of speakers to begin behaving like one another in conversations. Research has found that speakers entrain to both human and computer conversational partners. Evidence of entrainment has been found in multiple aspects of speech, including acoustic-prosodic, lexical, syntactic, and gestural. More interestingly, the strength of entrainment has been shown to be associated with numerous conversational qualities, such as naturalness of speech, task success, dialogue success, or social variables, such as cohesiveness. These two characteristics, commonness and association with positive conversational qualities, make entrainment an interesting research area for multiple disciplines, such as natural language processing, linguistics, and psychology.

Consider the small excerpt of a real conversation from the Teams corpus (Litman et al., 2016) between three speakers A, B, and C in Table 1. The speakers are playing a board game, *Forbidden Island*. In this game, each tile of the board can have three states: normal, flooded, and sunk. If a tile is flooded the players can unflood it with one of their actions if their pawn is on the flooded tile. This action is called unflooding. If a tile is sunk, it is removed from the board and this cannot be undone anymore. In the example in Table 1, speaker B in utterance number 2 uses the word "unsink" mistakenly instead of "unflood". Shortly after, speaker A also uses the same word. They continue using this word instead of "unflood" in this game. This is an example of lexical entrainment where speakers entrain on using the word "unsink".

[1] Other terms in the literature include accommodation, adaptation, alignment, convergence, coordination, and priming.

U_{C1} :	Ok so yeah that's my turn. Alright.
U_{C2} :	Um.
U_{C3} :	Mmhmm.
U_{B1} :	You should probably
U_{B2} :	**Unsink** the two things that have like
U_{C4} :	Yeah 2 Things on them
U_{C5} :	Um so let me think
U_{A1} :	You should yes, you should **unsink** your own thing.

Table 1: An example of lexical entrainment in a real conversation.

One of the interesting applications of entrainment research is spoken dialogue systems. Interactive conversational agents have been implemented in many devices including cars, phones, and TVs. Implementing dialogue systems with human-like characteristics is an open research problem. One of the most common human characteristics in conversation is entrainment. The first step towards this end goal of implementing dialogue systems with entrainment capabilities is to understand human entrainment behavior and to present measures to estimate it.

The development of methods for automatically quantifying entrainment in text and speech data is an active research area. While most research in this area has focused on quantifying the amount of entrainment between pairs of speakers, recent studies have started to develop measures for quantifying entrainment between larger groups of speakers (Friedberg et al., 2012; Danescu-Niculescu-Mizil et al., 2012; Gonzales et al., 2010; Doyle and Frank, 2016). Studying entrainment in multi-party environments is a topic of interest as teams, rather than individuals, are now the usual generators of scientific knowledge. How to optimize team interactions is a passionately pursued topic across several disciplines. Moreover, entrainment capabilities are useful for conversational agents that monitor and facilitate group interactions.

To date, mainly simple methods such as unweighted averaging have been used to move from pairs to groups, and the focus of prior work has been on text rather than speech (e.g., Wikipedia, Twitter, online discussion forums, and corporate emails). The focus of this research, unlike previous studies, is **multi-party spoken dialogue**. The goal of this

work is to develop, validate, and evaluate multi-party entrainment measures that can 1) be computed using language technologies, 2) **incorporate characteristics of multi-party interactions** and 3) **be associated with measures of team outcomes**.

1.2 THESIS OVERVIEW

To develop computational measures of multi-party entrainment, I study various dimensions described below:

Features: Research has been done on measuring entrainment on features from different linguistic levels. But, the association of entrainment on these features has generally not been investigated. I utilize features from two modalities available in spoken dialogue data (acoustic-prosodic and lexical), calculate entrainment using the same measures on them, and explore their relations and their utility to predict group outcomes. My experiments using existing measures show that there is a correlation between different modalities and a multimodal prediction model outperforms a unimodal model at the downstream task of predicting team outcomes.

Time Frame: Entrainment can be both a linear and dynamic phenomenon (De Looze et al., 2014). Feature similarity may increase somewhat linearly over the course of conversation (global). However, entrainment can also be short-term or local (e.g. only between adjacent turns), with feature similarity fluctuating within conversations. Speakers might show different entrainment behaviors. While one might entrain globally, another speaker might entrain locally. So, it is important to study these various aspects of entrainment, and this research develops local and global measures of entrainment in multi-party spoken dialogues. Developing measures of short-term (local) entrainment is more difficult to investigate in teams rather than dyads. While research on short-term pairwise entrainment typically examines similarity between two consecutive utterance or turns across speakers (Porzel et al., 2006; Levitan and Hirschberg, 2011; Danescu-Niculescu-Mizil et al., 2011; Mitchell et al., 2012; Ward and Litman, 2007b; De Looze et al., 2014), in multi-party conversations, dyad exchanges are not always adjacent, and/or participants can speak to two or more people at

the same time. Rather than attempt a complex manual annotation of addressee detection or solve an open algorithmic problem, I utilize a simple temporal window-based method. I hypothesize that there might be a time-lag between the source and the target utterances. The source and target utterances are a set of dyadic exchanges where the source utterance is uttered preceding to the target utterance and the target utterance could entrain to the source. So, using the window approach, I study the effect of time-lag on multi-party local entrainment. I did not find any significant result supporting this hypothesis.

From Pairs to Groups: To date, no matter what measure is used to estimate entrainment, mainly simple methods such as unweighted averaging have been used to move from pairs to groups. But, multi-party interactions are more complicated. Unlike two-party conversations, there are many pairwise relations in a group. Not all of these pairs show the same behavior. For example, a group might have both converging and diverging pairs. A speaker might converge to an interlocutor while diverging from another one. Moreover, using adaptive local entrainment measures, we are able to estimate not only the strength, but also the direction of pairwise entrainment. Directionality of the entrainment measure means that entrainment of speaker A towards B is not equal to the entrainment of speaker B towards A. Simply averaging pairwise entrainment values, as in the literature, ignores directionality and as a result ignores the dynamics of entrainment behaviors.

In this research, I propose global and local multi-party measures of entrainment that incorporate group entrainment dynamics. As for the global perspective, I propose a weighted convergence measure where the weights are defined based on the entrainment behaviors in the groups with the goal of decreasing the influence of outlier behaviors. Regarding the short-term (local) entrainment, I introduce graph-based vector representations of entrainment in multi-party groups. I propose to utilize pairwise directional measures of entrainment to represent multi-party entrainment as graphs. Then, I propose to utilize graph algorithms and represent nodes and graphs as vectors to encode the strength and the dynamics of entrainment relations. Finally, I propose approaches to learn the vector representations of the graphs (groups). The objective of the proposed vector representation is to encode the dynamics of entrainment behavior for each group and to achieve a more informative representation than a single-valued measure. The experiments indicate the promising performance

and utility of the proposed methods.

Individual and Team-level Attributes: Prior research has found relationships between non-conversational characteristics, such as gender, role, and expertise, and both the presence and utility of entrainment in dyads (Levitan and Hirschberg, 2011; Pardo, 2006) and teams (Yu and Litman, 2019). Given this knowledge, the research question is whether incorporating non-conversational characteristics (I only employ gender-composition in this study) enhances measures of multi-party entrainment. The proposed multi-task vector representation learning using an auxiliary task to predict gender-composition shows some promising results and encourages further investigation of this path.

1.3 RESEARCH QUESTIONS

I attempt to answer five main research questions:

- What is the relation between entrainment at different linguistic levels of acoustic-prosodic and lexical?
- Is there any time-lag between source and target utterances when measuring multi-party local entrainment?
- How do we incorporate the dynamics of local and global entrainment behaviors to introduce enhanced multi-party entrainment measures?
- How do we incorporate non-conversational factors such as gender-composition to enhance multi-party entrainment measures?
- Do the introduced measures have advanced utility at the downstream task of predicting team outcomes?

1.4 HYPOTHESES

First, I use a simple averaging method to extend global measures of proximity and convergence from pairs to groups. I explore the relation of acoustic-prosodic and lexical entrain-

ment. I hypothesize that:

- **H1***: Entrainment occurs at multiple linguistic modalities during conversation.*
- **H2***: Teams that entrain on one modality are more likely to entrain on the other, and vice versa.*
- **H3***: To predict team outcomes, a multi-modal model with entrainment features from acoustic-prosodic and lexical levels outperforms a unimodal model.*

Second, I propose a weighted multi-party convergence measure that incorporates the group dynamics. I hypothesize that:

- **H4***: A properly weighted multi-party convergence measure has a higher validity than the simple-averaging convergence.*
- **H5***: A properly weighted multi-party convergence measure has a higher utility than the simple-averaging convergence at the downstream task of predicting team outcomes.*

Third, for the short-term (local) entrainment, using the temporal window approach, I study the effect of time-lag between source and target utterances in multi-party entrainment. I hypothesize that:

- **H6***: Time-lagged multi-party entrainment estimated by the proposed window-based approach has a higher validity than the immediately preceding baseline.*
- **H7***: Time-lagged multi-party entrainment estimated by the proposed window-based approach has a higher utility than the immediately preceding baseline at the downstream task of predicting team outcomes.*

Fourth, I propose entrainment vector representation, a novel approach to encode both the strength and the dynamics of local entrainment behaviors in multi-party conversations. I hypothesize that [2]:

- **H8***: The proposed vector representation of multi-party entrainment outperforms the state-of-the-art adaptive entrainment baselines at the downstream task of predicting team outcomes.*

[2]The vectors are trained to have high validity. So, the validation experiment is not required.

Finally, I propose two approaches to enhance multi-party entrainment measures by incorporating team-level non-conversational attributes, such as gender-composition. I hypothesize that:

- **H9***: The proposed gender-aware extension of the entrainment measures outperforms the none-extended versions at the downstream task of predicting team outcomes.*

1.5 CONTRIBUTIONS

This thesis advances research on multi-party entrainment. It has multiple contributions. First, the focus of this research, unlike previous studies, is multi-party spoken dialogue. Second, I develop, validate, and evaluate novel multi-party entrainment measures that incorporate characteristics of multi-party interactions and are associated with measures of team outcomes. For the global entrainment, we present a weighted convergence measure. For the local entrainment, we present a graph-based entrainment vector representation. Moreover, this study provides insights on the relationship between two modalities of spoken dialogues (acoustic-prosodic and lexical) and whether entrainment occurs at multiple levels during conversations. Finally, I propose a novel approach to incorporate a team-level factor of gender-composition to enhance multi-party entrainment measures. All of the proposed works are in the direction of enhancing multi-party entrainment measures with the focus on spoken dialogues although they can also be employed on text-based communications.

1.6 THESIS OUTLINE

This chapter introduces entrainment, motivations for this study, research questions, hypotheses, and a summary of main contributions.

In chapter 2, I introduce the theory behind entrainment studies, discuss the related work on entrainment from different perspectives in detail, and explain where this work stands compared to the previous work.

In chapter 3, I introduce the Teams corpus which is used in this thesis and includes multi-party task-oriented spoken dialogues. I also describe the general setting of the experiments throughout this thesis, including the features, the pre-processing of data, and evaluation methods.

In chapter 4, I study the relationships of entrainment at the two linguistic levels, acoustic-prosodic and lexical, in multi-party dialogues (Rahimi et al., 2017). While there is considerable evidence for both exical and acoustic-prosodic entrainment, little work has been conducted to investigate the relationship between these two different modalities using the same measures in the same dialogues, specifically in multi-party dialogue. I measure lexical and acoustic-prosodic entrainment for multi-party teams to explore whether entrainment occurs at multiple levels during conversations and to provide insights on the relationship between these two modalities.

In chapter 5, I present a new weighting method to extend a dyad-level measure of convergence to a multi-party measure by considering group dynamics instead of simply averaging (Rahimi and Litman, 2018). I argue that although simple averaging is a good starting point to measure team entrainment, it has several weaknesses in terms of capturing team-specific behaviors specifically related to convergence (Rahimi et al., 2019). Experiments indicate validity and utility of the proposed weighted measure and also show that, in general, a proper weighting of the dyad-level measures performs better than non-weighted averaging in multiple tasks.

In chapter 6, I first discuss the limitations of existing short-term (local) multi-party entrainment measures, such as ignoring time-lag between source and target utterances, and ignoring the dynamics of entrainment relations in groups. To address such limitations, in the first part of this chapter (section 6.1), I investigate the time lag of influence in groups using an asymmetric local measure and the proposed temporal window algorithm. I found no evidence that considering time-lag improves the adaptive local entrainment measure. In the second part (section 6.2), I propose a novel graph-based vector representation for multi-party entrainment by incorporating the dynamics of entrainment relations. The proposed entrainment vector representation shows promising results compared to the state of the art. Finally, in chapter 7, I propose new methods to utilize non-conversational factors such as

gender-composition to enhance entrainment measures. The proposed multi-task learning where predicting gender-composition is an auxiliary task, shows significant improvements at predicting the favorable outcome.

Finally, in chapter 8, I summarize the major contributions and results of this work and present the limitations and possible future directions.

2.0 RELATED WORK

In the literature, linguistic entrainment is referred to with different terms, such as alignment, accommodation, and adaptation. Brennan and Clark (1996) define linguistic entrainment as the phenomena of speakers repeatedly using the same or closely similar terms to refer to the same objects in a conversation. They observed lexical entrainment where within-pair similarities on referral terms were significantly higher than between-pair similarities. Giles et al. (1991) defines speech accommodation as changes that people make to attune their communication to their speaking partners. The accommodation theory focuses on a specific form of entrainment, convergence, which is the increased linguistic similarity in the course of conversation. I, similar to the numerous studies of linguistic entrainment (Brennan, 1996; Nenkova et al., 2008; Levitan and Hirschberg, 2011; Doyle and Frank, 2016), refer to entrainment as a general term which covers different behaviors, such as similarity, convergence, synchrony, and proximity.

To explain the entrainment phenomena, Brennan and Clark (1996) emphasize more on automatic priming mechanisms of entrainment. They argue that conceptual pacts are established by speaking partners jointly and during the course of conversation. Also, speakers show different entrainment behaviors with different partners. This study emphasizes on recency, frequency of past references, informativeness and availability to explain how conceptual pacts are formed. In contrast, Accommodation Theory (Giles et al., 1991) emphasizes more on communicative and strategic nature of alignment. For example, it argues that "since speech is a way to express group membership, people adopt convergence or divergence in communication to signal a salient group distinctiveness, so as to reinforce a social identity".

2.1 ENTRAINMENT AT DIFFERENT LINGUISTIC LEVELS

Entrainment has been investigated at different linguistic levels. Empirical evidence of acoustic-prosodic entrainment has been reported in the literature. Matarazzo and Wiens (1967) showed that interviewers can influence latency or average duration of silence in interviewees' speech. In another study, manipulation in vocal intensity of the interviewers resulted in convergence of subjects' intensity to the new loudness levels (Natale, 1975). Guzzo et al. (1993) found that similarity of the speakers in terms of pitch and intensity was greater in the real conversations than in the simulated ones. Similarly, Ward and Litman (2007b) found entrainment on pitch and intensity as good measures to discriminate real and randomized conversations. At the phonetic level, Pardo (2006) found that speakers in conversational settings are susceptible to phonetic convergence. Levitan et al. (2011) investigated convergence on backchannel-preceding cues (BPCs) and found it to be significantly correlated with speech followed by backchannels. The results showed that speaking partners not only tend to use similar sets of BPCs, but also this similarity increases over the course of a dialogue. Levitan et al. (2012) investigated proximity and showed that conversational partners are more similar to each other than non partners in terms of acoustic-prosodic features, such as intensity mean, intensity maximum, and syllables per second. (Levitan and Hirschberg, 2011) and Lubold and Pon-Barry (2014) found evidence of entrainment on four acoustic-prosodic features (intensity, pitch, voice quality, and speaking rate) using three measures of proximity, synchrony, and convergence. Having children interact with a virtual human in a controlled experiment, Finkelstein et al. (2012) found that dialect of the virtual partner affects the prosody of young African American children.

Entrainment at the lexical level has been investigated extensively. Brennan and Clark (1996), Brennan (1996), and Metzing and Brennan (2003) studied entrainment on referring expressions. It has been shown that conversational partners entrain on referring expressions, they reach a conceptual pact, and continue to appeal to it in later references. Evidence of linguistic style matching was found in computer chats, twitter conversations, and spoken conversations (Niederhoffer and Pennebaker, 2002; Danescu-Niculescu-Mizil et al., 2011; Gonzales et al., 2010). Linguistic style refers to the components of language that are unre-

lated to the content. Linguistic Inquiry Word Count (LIWC) (Pennebaker et al., 2001) is applied to measure word usage in categories, such as articles, auxiliary verbs, and positive emotions. Linguistic style matching measures the similarity of speaking partners in using these word categories. In another study, Beňuš et al. (2012) showed evidence of entrainment on the use of filled pauses, such as *uh* and *um* in spoken dialogues. (Nenkova et al., 2008) examined entrainment on high-frequency words (the most common words in the corpus) and showed its correlation with the perceived naturalness of the dialogues and task success. Friedberg et al. (2012) found that there is a significant difference between lexical entrainment on project-related words of the high performing and the low performing student engineering groups.

Evidence of entainment in other levels, such as syntax (Branigan et al., 2000), pragmatic (Roche et al., 2012), and gesture (Alderisio et al., 2017; Richardson et al., 2009) has been found in the literature. Branigan et al. (2000) showed that the syntactic structure of utterances is influenced by interlocutors' previous utterances. Reitter and Moore (2006) observed that in task-oriented dialogues, speakers are more likely to use a syntax rule soon after their interlocutors used the same rule.

In this research, I focus on two linguistic levels aucostic-prosodic and lexical. In terms of acoustic-prosodic features, this research is built upon the work of Levitan and Hirschberg (2011). In terms of lexical features, our work is similar to (Nenkova et al., 2008; Friedberg et al., 2012; Niederhoffer and Pennebaker, 2002; Danescu-Niculescu-Mizil et al., 2011; Gonzales et al., 2010) where I measure entrainment on high frequency words, game-related words, and LIWC categories. Although, entrainment has been investigated exhaustively at different linguistic levels, the association between these levels are not investigated. In the first part of this work, I investigate the association between lexical and acoustic-prosodic entrainment in multi-party spoken conversations (Rahimi et al., 2017).

2.2 ENTRAINMENT MEASURES

Numerous entrainment measures are defined in the literature. Each of these measures focus on a different aspect of entrainment. Proximity (Levitan and Hirschberg, 2011) compares the similarity of speaking partners versus non-partners. Higher values indicate that speaking partenrs' similarity is more than random. Convergence (Natale, 1975; Street Jr, 1984; Coulston et al., 2002; Pardo, 2006; Ward and Litman, 2007a; Levitan and Hirschberg, 2011) measures the increase or decrease in the similarity of partners over time. Several studies (Brennan and Clark, 1996; Niederhoffer and Pennebaker, 2002; Pennebaker et al., 2001; Nenkova et al., 2008; Friedberg et al., 2012) have looked at the global similarities between speakers, while others (Lee et al., 2011; Danescu-Niculescu-Mizil et al., 2011; Fusaroli et al., 2012; Wang et al., 2015) have looked at conversational turn-by-turn (local) similarities. Truong and Heylen (2012) investigated dynamics of alignment over time using a windowed cross-correlation procedure. Their results showed weak tendencies toward synchrony. Measuring local *adaptive* entrainment, Danescu-Niculescu-Mizil et al. (2011) used a probabilistic framework; while Doyle et al. (2016b) estimated alignment using a generative model which provides a robust framework for sparse data. Jain et al. (2012a) used dynamic Bayesian networks to estimate alignment state at each conversational turn.

The scope of this research covers both global and local entrainment since different speakers might show different entrainment behaviors and it is important to define proper measures to estimate these different types of entrainment. In this work, I adapt several measures from literature including proximity and convergence (Levitan and Hirschberg, 2011) and adaptive entrainment measures (Danescu-Niculescu-Mizil et al., 2012; Doyle and Frank, 2016). These measures are either developed for pairwise interactions or leverage simple averaging to extend pairwise entrainment to multi-party entrainment. In this research, I propose novel approaches for incorporating the dynamics of entrainment behavior into measures of multi-party entrainment (Rahimi and Litman, 2018; Rahimi et al., 2019).

2.3 RELATION WITH SOCIAL OUTCOMES

Motivated by the theories relating entrainment to pragmatic goals (Giles et al., 1991), several studies have explored the relationship between entrainment and numerous social outcomes. Niederhoffer and Pennebaker (2002) found associations between Linguistic Style Matching (LSM) and perceived interaction quality (rapport) reported by the participants or the external judges. In contrast, Chartrand and Bargh (1999) did not find any correlations. They proposed a coordination-engagement hypothesis as an alternative to the coordination-rapport hypothesis: "The more that two people in a conversation are actively engaged with one another, in a positive or negative way, the more verbal and nonverbal coordination we expect." Lubold and Pon-Barry (2014) found a positive correlation between turn-by-turn local acoustic-prosodic entrainment and rapport. Results suggested that speakers entrain on pitch most significantly among other features when rapport is present. Michalsky et al. (2018) looked at the relation between prosodic entrainment and the perceived conversational quality in dating conversations. They found that conversational quality has a significant effect on a speaker's pitch entrainment. Pitch entrainment has also found to have a significant effect on perceived conversational quality. Natale (1975) looked at the vocal intensity convergence and showed that mean intensity convergence is positively correlated with the self-reported social desirability score of the individuals.

Levitan et al. (2018) investigated the relation between acoustic-prosodic entrainment and deception or trust. The results suggested significantly higher local entrainment on max intensity and jitter in conversations with deceptive speech than truthful speech, as well as higher local entrainment on mean intensity in conversations judged as deceptive than judged as truthful. Beňuš et al. (2018) investigated the link between subjects' trust toward an avatar, "expressed as the tendency to follow its advice", and acoustic-prosodic convergence or divergence of the avatar's voice towards the subjects. They found that females trusted the avatars who locally diverged on mean pitch, intensity, and speech rate. Reichel et al. (2018) measured prosodic entrainment in cooperative games. The investigation of interaction between gender and speaker role (describer vs. follower) showed that, in cooperative interactions, strategies are gender-dependent. For example, they found that female describers

entrain most and male describers least. Levitan et al. (2012) investigated the interaction between acoustic-prosodic entrainment, gender, and social behaviors. The results suggested significantly different entrainment behaviors for different genders. They found that male-female pairs entrained on more features than male-male pairs. Also, entrainment had the most significant correlations with the perception of social behaviors of mixed-gender pairs, and to the efficiency and flow of male-male pairs.

Xu et al. (2018) applied a generalized linear model to investigate the interaction between lexical entrainment and social power. The results suggested that entrainment is influenced by low-level linguistic features such as utterance length rather than social power. They suggest that the relation between social power and lexical alignment in (Danescu-Niculescu-Mizil et al., 2012) is not reliable since it did not control for low-level linguistic features. But, Beňuš et al. (2014) found similar results to (Danescu-Niculescu-Mizil et al., 2012) for acoustic-prosodic entrainment where lawyers entrain more to justices than justices to lawyers with respect to local entrainment on intensity.

Investigating the relation between entrainment and task performance, Nenkova et al. (2008) found a significant correlation between lexical entrainment (on most frequent words) and naturalness of dialogue; and also between entrainment on classes of common words and task success. Gonzales et al. (2010) found a positive relation between linguistic style matching and cohesiveness of groups in both face-to-face and computer assisted text-based communications, and a positive correlation with task performance in face-to-face environments. Fusaroli et al. (2012) also found that the alignment of particular task-relevant vocabularies strongly correlate with collective task performance. In the study of the relation between entrainment of students to a tutoring dialogue system, and learning rate, Thomason et al. (2013) found that male mean entrainment was significantly higher than female mean entrainment on min and max loudness features. They also found positive correlations of learning gain with entrainment for several pitch features. In another study, Friedberg et al. (2012) investigated lexical entrainment in engineering study groups and found that there is a significant difference between lexical entrainment on project-related words of the high performing groups, which tended to increase with time, and entrainment of the low performing groups, which tended to decrease with time.

Investigating local PCA-based vocal entrainment in relation to psychologist's affective rating of real couples, Lee et al. (2011) found that when a spouse is rated as having positive emotions, he/she has a higher value of vocal entrainment compared to when rated as having negative emotions. Analyzing lexical and syntactic alignment in peer-to-peer online support communities, Wang et al. (2017) and Wang et al. (2015) found that lexical alignment correlates with stronger emotional support while syntactic alignment positively correlates with informational support.

Given the evidence of the relation between entrainment and social outcomes and interaction qualities, as in the previous studies, I examine the relation between multi-party entrainment and the self-reported social outcomes as an extrinsic method to evaluate the quality entrainment measures. Our goal is to develop multi-party entrainment measures that are powerful at predicting social outcomes.

2.4 ENTRAINMENT AND DIALOGUE SYSTEMS

Several researches have tackled implementing dialogue systems with entrainment capabilities. Lubold et al. (2015) implemented a dialogue system that can dynamically adapt its pitch value to the user's mean pitch value by manipulating the text to speech output. They found a positive relation between perceptions of rapport and pitch adaptation. In another study, Hu et al. (2016) implemented a natural language generator that can entrain to "the user's referring expressions, tense-modality selection, verb and noun lexical selection, hedge and cue word choice, and syntactic template selection, or any combination of these". The results showed that "entraining on user's hedges increases perceptions of friendliness while reducing naturalness, while entraining on user's referring expressions, syntactic template selection, and tense/modal choices increase perceptions of both naturalness and friendliness." In an attempt to improve dialogue success in a spoken dialogue system, Lopes et al. (2015) proposed a lexical entraining dialogue system. It is trained to follow the user's choice of prime word if the system performance is not negatively affected. Otherwise, the system proposes a new prime. Results showed this entrainment strategy reduces out-of-vocabulary and

word error rate, and increases the number of correctly transferred concepts. Furthermore, Mizukami et al. (2016) found signs of entrainment on dialogue acts in a dialogue system, and also found that lexical entrainment is different depending on the dialogue act of the utterance. Sun and Morency (2011) proposed a new approach for dialogue act recognition by speaker adaptation which balance the influence of speaker specific and other speakers data in multi-party meetings. In another study, de Kok et al. (2013) presented a speaker-adaptive model for predicting listener responses. The speaker-adaptive model is created from a collection of dyadic interactions and is used to select the prediction model that reflects the characteristics of the speaker.

All these studies focus on pairwise dialogues. Implementing dialogue systems with entrainment capabilities for multi-party conversations has not been attempted. Although I do not attempt to implement a multi-party dialogue system in this research, this work is essential as the first step toward this goal as corpus studies are often done before the study of dialogue systems to understand the human's behavior and to develop proper measures.

2.5 MULTI-PARTY ENTRAINMENT

Entrainment studies have mainly focused on two-party conversations. Iwata and Watanabe (2013) have investigated entrainment in multi-party meetings although their main focus still remained on estimating influence relations for each pair of speakers using a probabilistic model and the expectation maximization (EM) algorithm. There have been several studies about entraiment in multi-party text-based communications, such as Twitter conversations and online health communities, or less spontaneous speaking conversations, such as supreme court hearings (Danescu-Niculescu-Mizil et al., 2011, 2012; Beňuš, 2014; Wang et al., 2014). Danescu-Niculescu-Mizil et al. (2012) developed a probabilistic measure to estimate turn by turn adaptive similarities of each individual to the group. They simply averaged these values to measure the group entrainment. These group entrainment values were used to compare justices and lawyers as two groups with different levels of social power. Beňuš et al. (2014) have also investigated entrainment in supreme court hearings between justice-lawyer

pairs using similarity based measures . Wang et al. (2014) studied multi-party conversations in online health communities by estimating word repetition between one or more prime posts and a target post. Doyle et al. (2016b) introduced a generative approach to measure entrainment using a hierarchical alignment model (Shin and Doyle, 2018; Doyle and Frank, 2016; Doyle et al., 2016a). The group layer in the hierarchy has latent variables to model the entrainment of each group. Although this measure is more robust at presence of sparseness, it is applied to measure entrainment in large groups such as Twitter conversations or email communication in corporations. But, it needs to be examined for the case of small groups where there is not much data available for each latent variable.

Face-to-face multi-party meetings have also been studied in several work. Hung et al. (2011) focused on speaker diarization where there is only a single audio source available. Oertel and Salvi (2013) investigated group involvement in eight-party face-to-face conversations. They define several group features based on gaze, such as presence, the fraction of subjects looking at other subjects to predict the group involvement category. Matsuyama and Kobayashi (2015) presented a computational model of conversation facilitation in small groups using a conversational robot. They proposed models to control and balance engagement and associate language generation with user models.

Gonzales et al. (2010) studied alignment in small groups of four to sixth members in both face-to-face and text-based communications. This study measured the linguistic style matching (LSM) of functional words for each group member by comparing each person's language with the overall percentage of the remaining group members. Then, a simple average of all group member's LSM score resulted in the group entrainment values. Friedberg et al. (2012) studied entrainment in student engineering groups. It used a similarity measure introduced by Nenkova et al. (2008) to estimate pairwise entrainment and then used simple or weighted averaging based on ratio of words spoken by each member to reduce the importance of speakers that may have talked very little. The weighted average did not result in any improvement in the results.

In this work, I investigate several aspects of entrainment in multi-party speaking groups that have not been investigated in the prior work. The goal is to gain better understanding of group entrainment behaviors and to develop enhanced multi-party entrainment measures

that take into account the dynamics of interactions in the groups and are more predictive of group social outcomes. Although our focus is on face-to-face speaking conversations, there are no limitations to apply the proposed methods on other types of conversation such as text-based communications.

3.0 DATA, FEATURES, AND EXPERIMENT SETUPS

3.1 TEAMS CORPUS

The Teams corpus (Litman et al., 2016) consists of over 45 hours of cooperative task-oriented dialogues within groups of three or four speakers, where audio and video files were collected and transcribed using best practices for computational processing. The corpus was collected in a laboratory experiment[1]. The laboratory setting enabled high-quality audio and video capture, while the experimental study allowed to collect measures of team processes.[2]

3.1.1 Task

The task is the cooperative board game, Forbidden Island^{TM}, where players take on the roles of adventurers seeking treasures on an island before it is flooded. The cooperative task-oriented nature of the game requires players to communicate to achieve their goals (e.g., discussing cards and strategies in real time, see Figure 1 for an example excerpt from a dialogue or Appendices for a complete dialogue), lending itself directly to eliciting entrainment. Further, the game gives each player a different role [3] to achieve the team goals, as well as game-specific terminology, generalizing to real-world situations with teamwork (e.g. avia-

[1]This work has been published in (Litman et al., 2016). My contribution in this paper is the entrainment analysis section. Also, I participated in the segmentation and the transcription part of the data collection.

[2]A lab experiment involving a two-player game requiring spoken communication was similarly used to collect the Columbia Games Corpus of 12 spontaneous task-oriented dyadic conversations, which has been used in multiple studies of two-party entrainment (Levitan and Hirschberg, 2011; Levitan et al., 2012, 2011). The Team's corpus is approximately 5 times larger, includes speech from teams rather than from dyads, and relatedly includes new types of team-related meta-data. The corpus also contains both video and audio as our dialogues were face-to-face rather than restricted to voice.

[3]There are four different roles in this game: Engineer, Pilot, Messenger, and Explorer. Teams with three participants do not have the role of Explorer.

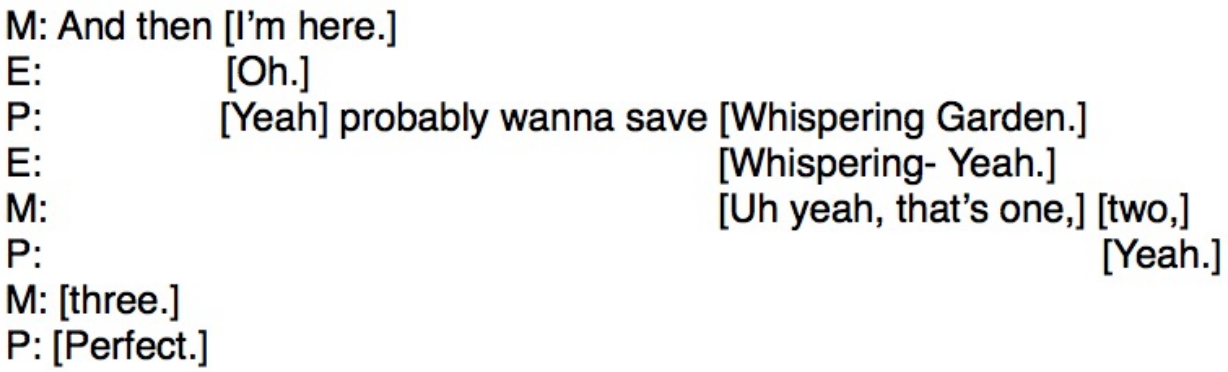


Figure 1: Dialogue excerpt from a Forbidden Island^{TM} game. E=Engineer, M=Messenger, and P=Pilot roles in the game. Square brackets indicate overlapping speech.

tion, health care). Two isomorphic versions of the game were constructed so that the first and second games would appear visually different but the difficulty level would be identical between and within teams. This isomorphism was accomplished by maintaining the position of tiles and cards that determined order-of-play and game difficulty, while systematically shifting the position of non-critical tiles and cards.

3.1.2 Recruitment

Participants aged 18 years and older who are native speakers of American English were recruited via electronic and hardcopy flyers and paid for their time. They were males and females of any ethnicity from a university and its surrounding community. To increase ethnicity, race, and age diversity (rare in corpora typically drawn only from student samples), advertisement was done in non-student locations in predominantly ethnic minority neighborhoods.

3.1.3 Procedure

As a team's participants arrived in the lab, each completed a questionnaire to collect personality, demographic, and other information such as experience with the game Forbidden Island^{TM}. Participants were then taught how to play the game by watching a video and playing a tutorial game, then given a few minutes to ask specific questions. Then each team played the game twice for no more than 35 minutes per game. Teams were told that not completing a game in 35 minutes counted as a loss, and that winning scores for the rest of the games would be inversely related to game length (a timer was displayed on a computer monitor during each game). Finally, both between and after the two games, all participants filled out questionnaires regarding their team processes.

3.1.4 Data Capture

Game participants were located around a round table 48 inches in diameter in the game-playing lab, enabling comfortable participant access to the game board. Each participant sat in a particular location depending on their role in the game. The survey data were collected in a separate workstation lab using Qualtrics, a web-based, survey software tool.

To collect high-quality speech data with minimal cross-talk, audio was recorded using Sennheiser ME 3-ew close-talk microphones. Each microphone was connected to a Presonus AudioBox 1818VSL multi-channel audio interface sampling at 96k, 24 bits. Audio recordings were monitored using Reaper Digital Audio Workstation v 4.76. Each game yielded one stereo recording with the synchronized speech from all speakers, along with 3 or 4 individual files (one per participant) representing the audio recording from each microphone. Reaper was used to render .WAV files with a 48000 Hz sampling rate and a 16 bit PCM Wav bit depth.

To complement the speech, four wall-mounted Zoom Q4 cameras captured WVGA/30 .MOV video recordings. The audio streams recorded from the cameras are at the central room, not the individual, level. A master audio signal was used to synchronize the videos with each other and with the audio from the microphones. Note that the videos also provide backup audio streams (recording at 256kbps AAC) for the microphones. In addition, the

	3-per.	4-per.
# of teams	36	27
avg g1 time	26.5	27.65
avg g2 time	18.1	18.7

Table 2: Team descriptives ($n = 63$).

videos provide information about the games that are not always obvious from the audio, as well as non-verbal data for future analysis (e.g., of gesture or posture).

3.1.5 Descriptive Statistics

The 216 participants in the experiment were on average 25.3 years old (min=18, max=67, *SD*=11.3). There were 135 females (62.5%) and 81 males (37.5%). The highest level of education (whether completed or not) ranged from high school (28 participants, 13.0%) to undergraduate (153 participants, 70.8%) to postgraduate/professional (35 participants, 16.2%). 145 participants (67.1%) were currently students. 35 participants (16.2%) knew at least one of their team members. The most frequent self-reported ethnicity/races were Caucasian (166), Asian (31), Black (24), and Hispanic (10) (multiple ethnicities were allowed).

Table 2 shows the distribution of the teams in the corpus by team size (3 versus 4 person). For each of these groups of teams, the table also shows the average time they took in minutes to play games 1 and 2, respectively. A 2-way ANOVA shows a significant within-team effect for game, with first games taking significantly longer than second games (27.1 vs. 18.4 minutes, $p < .001$). The average game length did not significantly differ by team size ($p > .3$).

One team's game 2 audio file was not properly saved during the experiment. To be consistent, I removed that team and used the remaining 62 teams in all of our experiments for both game 1 and 2.

3.1.6 Audio Segmentation and Transcription

After the experiment was completed, our multiple audio track speech was manually segmented and transcribed using the Higgins Annotation Tool[4]. To do transcription, each participant's speech is first segmented into inter-pausal units, pause-free chunks of speech from a single speaker (Levitan and Hirschberg, 2011). The threshold used for pause length (i.e., silence) for this corpus is 200 milliseconds. Once speech is segmented in a specific audio track, a corresponding text line appears where the transcriber manually types in the text for the corresponding audio segment. Five game 2 transcriptions were not completed, so I have transcriptions of 57 teams for game 2.

3.1.7 Questionnaire Data

The pre-game questionnaire was used to collect individual demographic information such as discussed in section 3.1.5, and self-reported data related to personality (John et al., 1991), cognitive styles (Miron et al., 2004), and collective orientation ("the propensity to work in a collective manner in team settings" (Driskell et al., 2010)). The between and post-game questionnaires elicited perceptions of team processes such as cohesion, satisfaction, and potency/efficacy (Wendt et al., 2009; Wageman et al., 2005; Guzzo et al., 1993). Such information is a novel resource for studying multi-party entrainment, since team processes have been shown to be positively related to performance (Beal et al., 2003; Mullen and Copper, 1994).

3.1.8 Perceived Interpersonal Team Outcome Measures

The between and post-game self-report questionnaires (see Appendices for the relevant subset of the questionnaires) that individuals took after each game include perceptions of cohesion (Wendt et al., 2009), general team satisfaction (Wageman et al., 2005), potency/ efficacy (Guzzo et al., 1993), and an adapted measure of shared cognition (Gevers et al., 2006), all of which had scale alpha reliabilities of $\geq .70$. Because these four measures are highly corre-

[4]http://www.speech.kth.se/hat/

lated, I z-scored each separate outcome and averaged these scores to make a single omnibus favorable group perception scale (alpha post-game1 = .78) and then averaged them for each team to create a team-level (positive) *favorable outcome*. The self-report surveys contained perceptions of three types of conflict (task, process, and relationship) (Jehn and Mannix, 2001) (each with alpha post-game1 $\geq$ .7). I z-scored the process conflict and averaged it in the groups to construct a team-level *process conflict outcome*. I chose process conflict over the other two conflict measures since process conflict was the only conflict measure that could be split at the median without making arbitrary choices [5].

3.2 FEATURES

In this thesis, I examine entrainment at acoustic-prosodic and lexical levels. At acoustic-prosodic level, I examine eight typical features that are used in the previous work (Levitan and Hirschberg, 2011; Levitan et al., 2012):

- Intensity mean
- Intensity max
- Intensity standard deviation
- Pitch (f0) mean
- Pitch (f0) max
- Pitch (f0) standard deviatio
- Jitter[6]
- Shimmer[7]

These features are computed from the audio files using Praat (Boersma and Heuven, 2002), an open-source audio processing software.

At the lexical level, I examine both content and functional words. In one setting, I measure entrainment on content words which are related to the topic of discussion and

[5]The median split is required for our classification tasks.

[6]The average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.

[7]The average absolute difference between consecutive periods, divided by the average amplitude.

are semantically meaningful. The detailed discussion is in Chapter 4. In another setting, for consistency with prior work (Danescu-Niculescu-Mizil et al., 2012), I measure lexical entrainment on eight LIWC-derived categories of function words (Pennebaker et al., 2001) that have little semantic meaning and are more relevant to style than content. These eight categories (451 lexemes total) are:

- Articles
- Auxiliary verbs
- Conjunctions
- High-frequency adverbs
- Impersonal pronouns
- Personal pronouns
- Prepositions
- Quantifiers

3.3 EXPERIMENT SETUPS

Here, I discuss the pre-processing and evaluation methods that I utilized throughout this work. The details specific to each experiment are discussed in the corresponding chapters.

3.3.1 Pre-processing

The raw audio is used to extract the acoustic-prosodic features. I did not remove the silences. Transcripts are pre-processed before extracting lexical terms by removing punctuation marks, converting all words to lower case, removing noises such as laughter, and removing any part of the transcript indicated as not fully understood by transcribers. For the study of content words, these steps are also added to the pre-processing: removing stop-words, removing filled pauses, and stemming the words.

3.3.2 Evaluation Approaches

To evaluate the proposed entrainment measures throughout this thesis, I will use two of the most common approaches. The first approach is to verify the validity of the proposed measures. The second approach is to test the utility of the measures at predicting social outcomes. Both of these methods are extrinsic evaluation. An intrinsic evaluation, if possible, is very expensive since human annotation of entrainment is a very complicated task specially at acoustic-prosodic level.

3.3.2.1 Evaluating Validation: One important criteria is that the entrainment measure should be valid or meaningful. In other words, the existence of entrainment should not be incidental. To evaluate this criteria, existing studies use fake vs real conversations (De Looze et al., 2014; Lee et al., 2011; Jain et al., 2012b). If the measure is valid, there should be a distinguishable difference between the real conversations and the fake conversations that are built using random permutations. Also, we might expect that a classification method that use entrainment values as features should be able to detect fake and real conversations better than random.

To evaluate validity of measures, I generate a random permutation of each game in the corpus. Then, I perform a binary classification using the team entrainment values as features to predict if each instance in data is real or permuted. So, the size of the data in this experiment is double the size of the real data.

To generate the permuted version of each game, from audio files or transcriptions, I randomly permute speech and silence intervals of each speaker independently. So, the speakers are the same in both fake and real versions of each team. As a result, the vocabulary of speakers are the same for both fake and real versions of each game. The only difference is the ordering of the speech and silence intervals for each speaker. Using these permutations, I generate the permuted version of each audio and transcription file for each game and I add them to the corpus.

3.3.2.2 Evaluating Utility at Predicting Team Outcomes: Another extrinsic evaluation approach that has been utilized to examine the quality of entrainment measures is to test their ability to predict speakers' or interactions' characteristics that are known to be associated to entrainment such as social statues, power, and positive or negative affect (Danescu-Niculescu-Mizil et al., 2011; Doyle et al., 2016b; Lee et al., 2011).

Similarly, I examine the relation between multi-party entrainment and the two self-reported measures of perception of team outcomes explained in section 3.1.8: *favorable outcome* and *process conflict*. I examine these relations using regression or classification tasks.

Regression analysis is done using the team outcome as the dependent variable and the multi-party entrainment as the independent variable. In classification task, I split the favorable outcome and process conflict at the median and generate two binary outcome variables. Then, I examine the utility of multi-party entrainment measures at predicting these binary outcome variables using binary classification tasks.

4.0 MULTI-PARTY ENTRAINMENT AT MULTIPLE LINGUISTIC LEVELS

Proximity and convergence are two important entrainment measures that has been used frequently in dyadic interactions (Levitan and Hirschberg, 2011; Levitan et al., 2012) and showed correlation with task and dialogue quality measures. The first step toward the analysis of entrainment at multi-party dialogue is to extend the dyadic measures to multi-party ones and show that entrainment exist in the corpus.

Some research has attempted to show that entrainment co-occurs at multiple linguistic levels, with varying success (Reitter and Moore, 2007; Heath, 2017). Recent attempts to quantify entrainment in multi-party dialogues has mainly been conducted on written language from text based communication (Internet forums, Twitter, emails, etc.) (Danescu-Niculescu-Mizil et al., 2011; Mukherjee and Liu, 2012; Munson et al., 2014). Consequently, there are few studies that measure multi-party entrainment on spontaneous spoken conversations (Danescu-Niculescu-Mizil et al., 2012; Friedberg et al., 2012; Litman et al., 2016). Since spoken dialogues have several linguistic modalities, studying the relation between entrainment at different modalities is another interesting research question that has not get much attention yet. Further, because many multi-party entrainment studies measure semantically light words (like filled pauses or grammatical function words) (Gonzales et al., 2010; Beňuš et al., 2012), there are few investigations that analyze group entrainment on lexical items with more semantic content. In this chapter, I take a multi-modal approach by integrating and extending prior work on semantically rich lexical (Friedberg et al., 2012) and acoustic-prosodic group entrainment.

I extend the dyadic convergence and proximity measures to multi-party by simple averaging of pairwise entrainment as it was done in previous studies (Friedberg et al., 2012;

Gonzales et al., 2010). Next, I perform an analysis on acoustic-prosodic and lexical entrainment to examine three hypotheses: H1) both acoustic-prosodic and lexical entrainment occur in multi-party team dialogues, H2) teams that entrain on one linguistic level are more likely to entrain on the other level, and H3) a multi-modal model with lexical and acoustic-prosodic entrainment features will better predict team outcomes than a unimodal model. I believe that investigating these hypotheses will provide insights about entrainment at multiple modalities.

In the next section, I describe the entrainment measures, convergence and proximity, that I adopt from prior works (Levitan and Hirschberg, 2011; Friedberg et al., 2012; Litman et al., 2016). Then, I explain the methods that I use to extract acoustic-prosodic and lexical features for computing entrainment. Finally, I will explore the three proposed hypotheses and discuss the results.

4.1 MULTI-PARTY PROXIMITY AND CONVERGENCE

In this chapter, I utilize two conversation-level (global) measures of entrainment: *proximity* and *convergence*. Proximity measures the degree of similarity between members within a team relative to participants in other teams. Convergence measures the change in similarity of teammates over time. These two forms of entrainment are shown in Figure 2. I utilize multi-party proximity and convergence measures [1] which average the corresponding dyad-level measures (Levitan and Hirschberg, 2011).

I quantify average distance of team partners ($TDiff_p$) using Equation 4.1, where $|team|$ refers to the size of the corresponding team and $feature_i$ (e.g., from the acoustic-prosodic and lexical levels described below) is the value of the corresponding feature for speaker i:

$$TDiff_p = \frac{\sum_{\forall i \neq j \in team}(|feature_i - feature_j|)}{|team| * (|team| - 1)} \tag{4.1}$$

To quantify team distance to the set of non-team (i.e. other) participants $|X|$, I follow a

[1] This part of my research is originally published in (Litman et al., 2016).

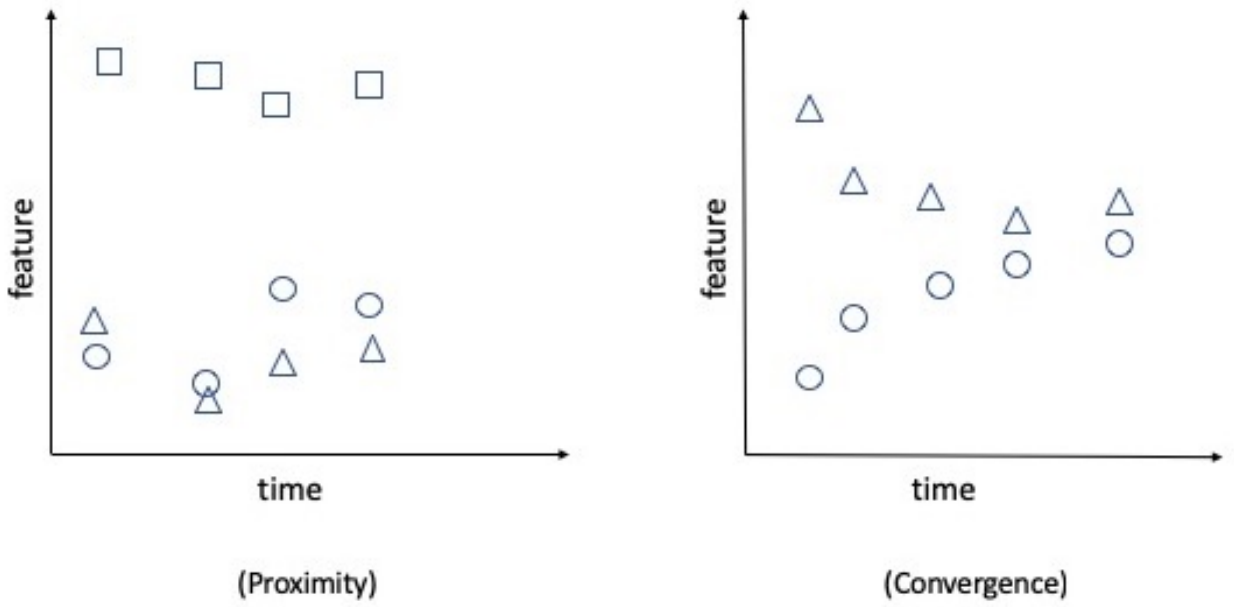


Figure 2: Proximity and Convergence. The circles and triangles represent feature values from two different speakers partnered in a conversation. Squares represent feature values of a speaker from another group (not partnered with the other two speakers).

similar calculation in Equation 4.2:

$$TDiff_o = \frac{\sum_{\forall i \in team}(\frac{\sum_{\forall j \in X} |feature_i - feature_j|}{|X|})}{|team|} \tag{4.2}$$

Proximity and convergence are then defined in Equations 4.3 and 4.4, respectively. Greater positive values indicate more team entrainment. For calculating convergence, note that the feature values are extracted from 2 different non-overlapping intervals.

$$proximity = TDiff_o - TDiff_p \tag{4.3}$$

$$convergence = TDiff_{p,Interval1} - TDiff_{p,Interval2} \tag{4.4}$$

4.2 ACOUSTIC-PROSODIC FEATURE LEVEL

Consistent with prior work on dyad entrainment (Levitan and Hirschberg, 2011; Lubold and Pon-Barry, 2014; Borrie et al., 2015), I extract features of pitch, intensity, jitter, and shimmer using Praat software (Boersma and Heuven, 2002). Pitch describes the frequency of sound waves, intensity describes the loudness or energy transported by the wave, and jitter and shimmer (acoustic characteristics of voice quality) measure frequency and energy variation, respectively. Specifically, I extract the following 8 features: maximum (max), mean, and standard deviation (SD) of pitch; max, mean, and SD of intensity; local jitter; and local shimmer. The features are extracted first from the whole conversation at game level, then from two non-overlapping intervals.

4.3 LEXICAL FEATURE LEVEL

Since (Nenkova et al., 2008; Friedberg et al., 2012) showed that speakers' entrainment on high-frequency words was significantly correlated with task success, social variables, etc., I

follow a similar methodology in this work. First, I select a set of lexical terms that I want to examine for entrainment. Second, I extract the frequencies for each term overall and for each speaker, with the latter normalized by the total number of terms uttered by the speaker. These normalized counts are the features in my entrainment measures. As in (Friedberg et al., 2012), I sum all term entrainment scores to get one final group entrainment score.

I first used existing methods to extract a set of terms:

HighTF-all: This term frequency method replicates (Nenkova et al., 2008). I choose the 25 most frequent words in the corpus. The idea behind this method is to avoid sparsity in my feature set.

HighTF-game: This method replicates (Friedberg et al., 2012). I choose the 25 most frequent words for each *individual game* (rather than all corpus games). The idea behind this and all other game-based methods is to select terms that are specific to each team.

ProjectWords-all: This method replicates (Friedberg et al., 2012). I extract game related words from the instructions that were handed to participants. Then, to avoid selecting rare words in the corpus, I select the 25 most frequent words in the corpus which occur in the instruction materials to avoid selecting off-topic words.

Since speakers may exhibit variation in the forms of words that they use, I hypothesize that extracting project words might not be the best approach for choosing game-related words. Consequently, I add a new data driven term-extraction approach utilizing topicSig [2], an automatic topic signature acquisition algorithm that recognizes relevant topic words in a document (Lin and Hovy, 2000). The algorithm identifies key terms by associating a certain topic with a signature, or a vector of related words and their associated weights of relation to the topic. Relational weights are calculated according to a likelihood ratio (λ) that compares the competing hypotheses values that the probability of the presence of a given word is indicative of a certain topic versus the probability that it is not. Probabilities are calculated by weighing the distribution of a certain word in the target corpus against its distribution in a background corpus. The transformed quantity $-2\log(\lambda)$ asymptotically follows a chi-square distribution, allowing us to select a significance level to set a threshold for selecting topic words. For my background corpus, I used part of transcripts for second

[2] topicSig is a Java implementation of Annie Louis' algorithm.

Method	Top 10 Terms
HighTF-all	yeah, ok, one, oh, two, move, card, right, get, three
ProjectWords-all	one, two, move, card, right, get, three, give, shore, turn
topicSig-all ($\lambda = 10$)	templ, treasur, card, discard, draw, gate, pilot, pile, flood, action

Table 3: Top 10 terms from different extraction methods.

games in the Teams corpus. Finally, I extract two sets of terms using the topic signature algorithm:

topicSig-all: I extract the topic signature when the target is the whole corpus. I select $\lambda = 10$ (p = 0.016), which is the default algorithm setting, and $\lambda = 4$ ($p = 0.05$) as this is a standard cutoff for statistical significance.

topicSig-game: I extract the topic signature for only the first game for each team. To address data sparsity issues, I only use the more lenient $\lambda = 4$ (p = 0.05) and only select terms that occur more than 5 or 10 times[3] for analysis.

To illustrate the terms selected by each type of lexical term selection method, Table 4.3 shows the top 10 terms (sorted by frequency and shown stemmed) for the "all" versions of each method. HighTF-all includes words such as "yeah", "oh", "ok" which are not semantically-rich. The ProjectWords-all terms do not have this issue since words absent from the project materials are omitted. The topic signature method also avoids this issue and additionally omits the most general project words.

[3]These cutoffs should be tuned to optimize performance in the future.

4.4 EXPERIMENTS AND RESULTS

4.4.1 Preprocessing

The raw audio is used to extract the acoustic-prosodic features. These features are normalized by gender when computing proximity since team partners ($TDiff_p$) are compared with participants from other teams ($TDiff_o$). The normalization is done using z-scores:

$$z = \frac{v_i - \mu}{\sigma} \tag{4.5}$$

where v_i is the value of the feature for speaker i, μ is gender mean value of the feature, and σ is gender standard deviation of the feature. I do not normalize for convergence since only team partners are compared ($TDiff_p$) (at different time intervals).

The transcripts are preprocessed before extracting lexical terms by: removing punctuation, converting all words to lower case, removing stop-words, removing filled pauses and noises such as laughter, removing any part of the transcript indicated as not fully understood by transcribers, and stemming all words.

Since convergence requires feature extraction from two non-overlapping intervals, in this section, I break each conversation into two equal halves. To be consistent, I use time to compute the halves for both the audio and transcripts. For transcripts, all utterances that begin before the breaking point are included in the first half, regardless of when they end, and the rest are included in the second half.

4.4.2 Experiment 1: Unimodal Entrainment

My first hypothesis is that both acoustic-prosodic and lexical entrainment exist in the corpus. To test this hypothesis, I measure entrainment (proximity and convergence) using the methods for computing lexical and acoustic-prosodic features and look for statistically significant results in the corpus. I employ the Student's paired t-test to determine if the differences between partners and others (proximity), and the differences between the first and second intervals (convergence) are significant.

Feature	**Proximity**	**Convergence**
Intensity max	3.327*	-0.189
Intensity mean	3.264*	0.413
Intensity SD	1.823+	-1.667
Pitch max	1.171	0.952
Pitch mean	0.550	-1.782+
Pitch SD	2.741*	1.319
Jitter	2.159*	2.234*
Shimmer	2.444*	2.748*
HighTF-game	-0.879	-0.102
HighTF-all	3.919*	0.657
ProjectWords-all	2.545*	0.333
topicSig-game (>10)	1.156	-0.790
topicSig-game (>5)	0.381	-1.174
topicSig-all ($\lambda = 4$)	4.265*	0.644
topicSig-all ($\lambda = 10$)	0.606	0.647

Table 4: The T-statistic of paired t-test on proximity and convergence at acoustic-prosodic and lexical levels. The positive T-statistic is a sign of entrainment. The entrainment is significant if the p-value is < 0.05 (*) and trending if < 0.1 (+).

The results are in Table 4. At the lexical level, HighTf-all, ProjectWords-all, and TopicSig-all ($\lambda = 4$) show significant proximity. All features except HighTF-game show proximity (have positive T-statistic) which means the partners are more similar to each other than to non-partners. At the prosodic level, only pitch-mean and pitch-max do not show significant proximity. These results indicate that proximity occurs at both acoustic-prosodic and lexical levels. However, I only see significant lexical proximity when the lexical terms are extracted from the whole corpus (as opposed to from an individual game). Even though I filtered out very low frequency terms (below either 5 or 10) when using topicSig-game, data sparsity might still be an issue and needs further investigation.

As for the convergence measure, Pitch mean shows a near significant divergence and Jitter and Shimmer show significant convergence from the first to the second interval. None of the lexical features show any significant results although HighTF-all, ProjectWords-all, and TopicSig-all show positive convergence. Note that the number of observed significant entrainment results are reduced in both linguistic levels for convergence as compared to proximity. For the lexical level, this might be because of the short length of the dialogues. Because the term frequencies are not very high for each individual in each game, when I break the game into halves these numbers are even smaller in each interval.

As a validity check for convergence, for each of the 62 teams, I constructed artificial versions of the real conversations between team members: For each member of the team, I randomly permuted the silence and speech intervals extracted by Praat. For the lexical entrainment, I randomly permuted the speech and silence intervals based on transcriptions. Ideally, I should not see evidence of convergence within these constructed conversations. The results of paired student t-test, confirm that there is no significant entrainment on the fake games, for all acoustic-prosodic and lexical features.

In conclusion, my first hypothesis is supported. Both acoustic-prosodic and lexical entrainment exists in my corpus.

4.4.3 Experiment 2: Multimodal Co-Occurrence

My second hypothesis is that entrainment not only separately occurs at both acoustic-prosodic and lexical levels but also it co-occurs across levels for individual teams. In other words, the teams that show entrainment in one level are more likely to show entrainment in the other level. To investigate this hypothesis, I quantify the correlations between acoustic-prosodic and lexical entrainment on both proximity and convergence. Significant positive correlations will support this hypothesis.

To calculate correlation, I employ the Spearman correlation coefficient. The significant results are in Table 5. In terms of proximity, multiple acoustic measures have significant (Pitch mean and max) or trending (Intensity SD and max) positive correlations with at least one lexical measure. In terms of convergence, Intensity SD also has significant positive lexical correlations. As with Experiment 1, there are more results when proximity rather than convergence is used to measure entrainment. In contrast to Experiment 1, however, extracting terms at the game rather than the corpus level now seems to be more useful.

In conclusion, I observe that within a team, lexical and acoustic-prosodic entrainment show signs of positive correlation, which supports my second hypothesis. However, I also see some negative correlations (indicating inverse relationship between the two variables) which needs further investigation.

4.4.4 Experiment 3: Benefit of Multimodal Analysis

My third hypothesis is that a multimodal model which includes both lexical and acoustic-prosodic entrainment as predictors of team outcomes will outperform a unimodal acoustic-prosodic model. To test this hypothesis, I use a hierarchical multiple regression to first construct a model with only acoustic-prosodic entrainment features (Model 1), then add lexical entrainment features to construct a multimodal model (Model 2). Significant improvement in the second model will support my hypothesis.

Table 6 shows the hierarchical regression results for each of the *favorable* and *conflict* team outcome measures introduced in Section 3.1.8. For example, when all of the acoustic-prosodic entrainment values were entered as potential independent variables for predicting

	Lexical	Acoustic	r
Proximity	HighTF-all	Intensity SD	0.245^{+}
		Pitch mean	0.279^{*}
	HighTF-game	Pitch SD	-0.263^{*}
	topicSig-game (>5)	Pitch mean	0.270^{*}
		Pitch max	0.261^{*}
	topicSig-game (>10)	Pitch SD	-0.218^{+}
		Intensity max	0.227^{+}
Convergence	HighTF-all	Intensity SD	0.283^{+}
	HighTF-game	Intensity SD	0.292^{*}
	topicSig-game (>5)	Pitch SD	-0.255^{*}
		Intensity SD	0.272^{*}

Table 5: Spearman correlation (r) between lexical and acoustic-prosodic proximities and convergence. The correlation is significant if the p-value is < 0.05 (*) and trending if it is < 0.1 (+).

Favorable outcome, a significant model containing Pitch mean proximity and Intensity SD proximity resulted (Model 1). After lexical entrainment features were considered, topicSig-all convergence was added to create Model 2 (which still includes the Model 1 features). The standardized βs indicate the effect size and direction of the individual independent variables on the dependent variable, whereas the R^2 indicates the effect size of the model with all variables.

Dependent		**Features**	**M1 (β)**	**M2 (β)**
Favorable		P-Intensity SD	0.206+	0.249*
		P-Pitch mean	-0.343*	-0.301*
		C-topicSig-all		0.300*
	R^2		0.162	0.248
	F		5.695*	6.382*
Conflict		P-Pitch mean	0.439*	0.411*
		C-topicSig-all		-0.208+
	R^2		0.192	0.235
	F		14.299*	9.064*

Table 6: Regression between entrainment and favorable (conflict) outcome measures. C, P, M refer to convergence, proximity, model. Significant / trending results if p-value is < 0.05 (*) or < 0.1 (+).

For Favorable outcome as the dependent variable, the independent entrainment variables which are significant covariates and that appear in the final hierarchical model are Intensity SD proximity, pitch mean proximity, and topicSig-all ($\lambda = 10$) convergence. Two of these (pitch mean proximity and topicSig-all ($\lambda = 10$) convergence) also appear in the Conflict model. For multi-modal prediction of Favorable outcome, the amount of variance explained is significant above and beyond the variables entered in Model 1, $\Delta R^2 = 0.086$, $\Delta F(1, 58) = 6.662$, $p = 0.012$. There was a significant positive association between topicSig-all and Favorable outcome. For Conflict, the amount of variance explained by Model 2 over Model 1 is trending, $\Delta R^2 = 0.043$, $\Delta F(1, 58) = 3.284$, $p = 0.075$, with a significant negative association between topicSig-all convergence and Conflict. I performed a similar experiment by first entering the lexical entrainment features to the model and I got the same conclusion. These results support my last hypothesis that multimodal models predicting team outcomes using both lexical and acoustic-prosodic entrainment will outperform unimodal models considering only acoustic-prosodic entrainment.

4.5 CHAPTER SUMMARY

In this chapter, I examined existence of entrainment (proximity and convergence measures) on a multi-party spoken dialogue corpus on two linguistic levels of acoustic-prosodic and lexical. I used simple averaging to calculate team entrainments. I showed that entrainment exists at both lexical and acoustic-prosodic levels; entrainment at these levels has significant correlation with each other; and a multi-modal model using both levels can predict the team outcomes significantly better than a unimodal model.

5.0 MULTI-PARTY WEIGHTED CONVERGENCE

Recently, a few researchers have studied multi-party entrainment in online communities and conversation groups (Gonzales et al., 2010; Friedberg et al., 2012; Danescu-Niculescu-Mizil et al., 2012; Litman et al., 2016). Regardless of which approach these studies use to measure entrainment, they utilize simple averaging to extend pair-level measures to group-level ones, as I did in the previous chapter. But, the question is can I do better than simple averaging? Although we get valid results using simple averaging, we are discarding the properties and behaviors of different groups that might actually improve the accuracy of group entrainment measures.

In this chapter, first, I perform a case-study analysis on convergence measure and I show that simple averaging is unable to capture at least two properties of multi-party convergence. So, in order to improve the limitations, I propose a new weighted convergence measure that utilizes group dynamics.

5.1 NON-WEIGHTED CONVERGENCE FOR MULTI-PARTY DIALOGUE

The convergence measure that I extend in this study is adopted from prior work. Originally, convergence between dyads (Levitan and Hirschberg, 2011) was measured by calculating the difference between the dissimilarity of speakers in two non-overlapping time intervals. If the dissimilarity in the second interval was less than in the first, the pair was said to be converging.

Extending this work, multi-party convergence (Litman et al., 2016) was measured using Non-Weighted (NW) averaging of each pairs' convergence, as shown in Equations 5.1 and

5.2:

$$GroupDiff_t = \frac{\sum_{\forall i \neq j \in group}(|f_{i,t} - f_{j,t}|)}{|group| * (|group| - 1)} \tag{5.1}$$

$$Conv_{NW} = GroupDiff_{t_1} - GroupDiff_{t_2} \tag{5.2}$$

$GroupDiff_t$ corresponds to average group differences calculated for linguistic feature f in time interval t for all pairs (i,j) in the group. $|group|$ refers to the size of the group. The convergence is the difference between $GroupDiff$s in two intervals.

5.2 CASE STUDY ANALYSIS OF CONVERGENCE

I divide each game into four equal disjoint intervals to get better insight of the behavior of individuals, as opposed to comparing two intervals. I extract the following 8 acoustic-prosodic features: maximum (max), mean, and standard deviation (SD) of pitch; max, mean, and SD of intensity; local jitter[1]; and local shimmer[2]. The features are extracted from each of the four intervals of each speaker in each team.

First, I perform a significance test to find out which features show significant convergence and on which intervals. The results of the repeated measures of ANOVA with interval as a factor with 4 levels are shown in Table 7. 'c' and 'd' are indicative of significant convergence and divergence on the corresponding intervals respectively. For example, speakers are significantly converging on shimmer from interval 1 to 3.

5.2.1 Is Simple Averaging a Proper Approach?

While previous studies have averaged pairs' entrainment to measure multi-party entrainment, it remains uncertain whether simple averaging is an optimal approach. What are the flaws and weaknesses of this approach and how can we improve them?

[1]The average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.

[2]The average absolute difference between consecutive periods, divided by the average amplitude.

Features	ANOVA	Pairwise Comparisons					
		1-2	1-3	1-4	2-3	2-4	3-4
Pitch-max							
Pitch-mean	*			d		d	d
Pitch-sd							
Intensity-max							
Intensity-mean							
Intensity-sd							
Shimmer	*		c	c	c	c	
Jitter	*		c	c			

Table 7: The results of the repeated measures of ANOVA. * indicates the p-value < 0.05. Pairwise comparisons indicate which intervals are significantly different. The direction (convergence or divergence) is represented by c and d respectively.

I argue that there are some group-specific behaviors that are not properly quantified using the simple averaging method. To demonstrate this with real examples from the corpus, I perform a within-team analysis in which I examine the behavior of individuals within each team and the relationship between their behaviors and team convergence.

For this purpose, I draw the plots of raw values of the feature for each team on all 4 intervals in game 1. I chose jitter and shimmer as my features of interest since they are the only features that demonstrated significant convergence. I sort all the teams by their convergence values computed by Equation 5.2. For example, the convergence of each team from interval 1 to interval 4 is defined as $GroupDiff_1 - GroupDiff_4$.

I examined the plots of all diverging, converging, and maintaining teams. I argue that there are at least two general cases that the simple averaging approach is unable to capture in the groups' behavior. I describe these two cases as follows.

First, how many of the team members are required to converge in order to consider the

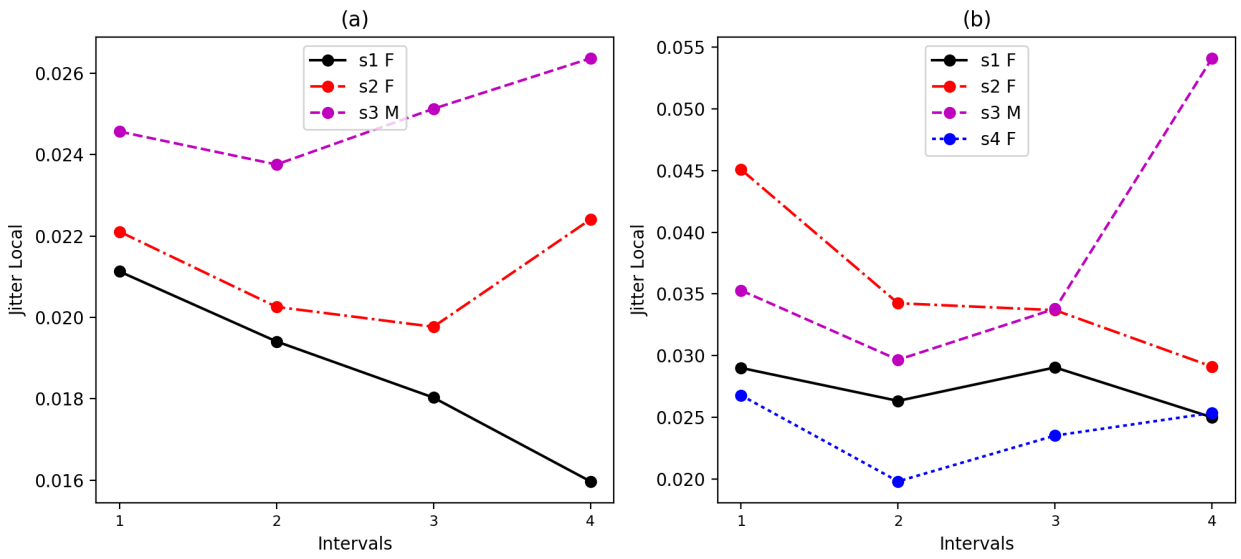


Figure 3: The plot of jitter of individuals over all four intervals in two diverging teams. Each point is the jitter value calculated for corresponding speaker at corresponding interval using the Praat software. S is short for speaker. M and F are indicative of gender. (a) Convergence from interval 1 to 4 calculated using Equation 5.2 is equal to −0.0139 (b) Convergence from interval 1 to 4 calculated using Equation 5.2 is equal to−0.0199

team to be converging overall? According to the simple averaging method, the answer is that the number of converging or diverging pairs does not matter. As long as the average convergence is higher than the average divergence, we consider the team to be converging. I argue that this answer is not accurate. For example, consider Figure 3. Each of the plots in this figure shows the values of jitter for each individual [3] in a team over the four intervals. Comparing the first and the last intervals, it appears that Figure 3(b) is the most diverging team in the corpus, based on the convergence measure. But, unlike the team in Figure 3(a), where all the speakers are diverging from each other, speaker 3 is the only participant to diverge from the team, while the rest of the speakers converge. The question is, how much

[3]I included the gender of speakers in the plots. But, there is no significant effect of gender composition of the teams on convergence value.

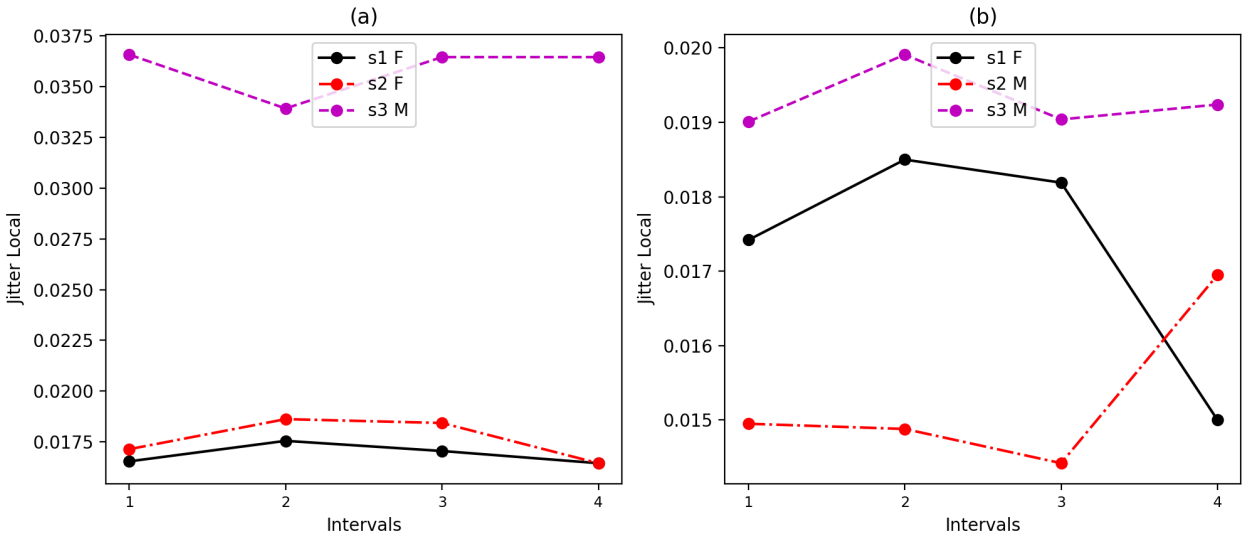


Figure 4: The plot jitter of individuals over all four intervals in the two most maintaining teams. Each point is the jitter value calculated for corresponding speaker at corresponding interval using the Praat software. S is short for speaker. M and F are indicative of gender. (a) Convergence from interval 1 to 4 calculated using Equation 5.2 is equal to 0.00006 (b) Convergence from interval 1 to 4 calculated using Equation 5.2 is equal to -0.00036

should speaker 3 influence the team convergence in this group?

I argue that the influence of a speaker, such as speaker 3, should lessen if his or her behavior is in the opposite direction of the team behavior. I hypothesize that the solution to this problem is to use a weighted average, where the weights are defined based on the number of speakers that have the same behavior in the team. For example, the weight of a diverging speaker should be the percentage of diverging individuals in the team. I explain this proposed approach in the next section.

Second, does the convergence or divergence of speakers in teams have an absolute meaning, similarly as in pairs? An individual might converge to one teammate while diverging from another one. How do these conflicting behaviors affect the team measure? For example, consider the two teams in Figure 4. Comparing the first and the last intervals, these two

teams have the closest convergence value to zero in my corpus, meaning they are the most maintaining teams. The plot in Figure 4(a) is an obvious case of maintenance where none of the team members change their feature values to converge toward or diverge away from the others. But, in Figure 4(b), speaker 1 changes her initial state to converge to speaker 2 while she diverges from speaker 3. I hypothesize that taking into account the self-difference, or how much each speaker's feature value has changed over time, will help to resolve this issue. This hypothesis is not examined in this work and needs to be investigated in future.

5.3 WEIGHTED CONVERGENCE FOR MULTI-PARTY DIALOGUE

In the next subsections, I introduce two weighted variations of convergence: a baseline based on participation ratios (Friedberg et al., 2012), and the proposed method based on group dynamics.

5.3.1 Baseline: Weighting Based on Participation

The idea behind this approach is that the weights for speakers that may have talked very little should be reduced. In prior work on multi-party lexical entrainment (Friedberg et al., 2012), speaker participation was measured by number of uttered words; the participation ratios of speaker pairs were then used as the weights.

For acoustic-prosodic features, I measure speaker participation by amount of speaking time. The Participation Ratio (PR) of each speaker in a given temporal interval is their total speech time divided by the duration of the interval including silences. Speech and silence periods are automatically annotated using Praat (Boersma and Heuven, 2002).

Convergence for pair $p = (i, j)$ and for two disjoint intervals t_1 and t_2 is calculated as in Equation 5.3:

$$Conv_{p=(i,j)} = (|f_{i,t_1} - f_{j,t_1}| - |f_{i,t_2} - f_{j,t_2}|) \tag{5.3}$$

The Participation-based Weighted (PW) convergence for a group is then computed as

the normalized weighted average of convergence of all pairs p in the group:

$$Conv_{PW} = \frac{\sum_{\forall p \in group}(Conv_p * PR_p)}{Num_p \sum_{\forall p \in group} PR_p} \tag{5.4}$$

Num_p indicates number of pairs, and Participation Ratio for a pair, PR_p, for the two intervals is the sum of PRs for both speakers and in both intervals.

5.3.2 Proposed: Weighting Based on Group Dynamics

Although participation-based weighting decreases the contribution of less active speakers when calculating group convergence, it does not take group convergence dynamics into account. I hypothesize that it might instead be better to decrease the contribution of speakers whose convergence behaviors differ from the rest of the group (e.g., $Speaker3$ in Figure 3(b)). To tackle this issue, I use weighting to decrease the contribution of outlier speakers. In particular, I propose that the weight for a speaker should be the percentage of individuals who have the same convergence behavior as the speaker.

Equation 5.5 defines the proposed Group Dynamic-Based Weighted (GDW) convergence measure:

$$Conv_{GDW} = \sum_{g \in G} \frac{|g|}{|N|} * \frac{\sum_{i \in g} \sum_{j \neq i \in N} Conv_{ij}}{|Num_{pair}|} \tag{5.5}$$

G is a set including three categories: $G = \{Converging, Diverging, MixedBehavior\}$, g is a set of all individuals who belong to a category in G, $|N|$ is the number of all speakers in the group, $Conv_{ij}$ is the convergance of the pair (i, j) calculated by Equation 5.3, and $|Num_{pair}|$ is the number of pairs.

Consider the example in Figure 3(b). There are 12 pairs (6 unique pairs since convergence is a symmetric measure). Each speaker is in three unique pairs with the other three members of the group.

Speaker1 pairs: (1,2),(1,3),(1,4)
Speaker2 pairs: (2,1),(2,3),(2,4)
Speaker3 pairs: (3,1),(3,2),(3,4)
Speaker4 pairs: (4,1),(4,2),(4,3)

If all conversational pairs that a speaker is involved in have positive convergence values, the speaker is converging to the group and has the *Converging* category. If all involved

pairs have negative value, the speaker is diverging from the group. Else, the speaker has a mixed-behavior.

The weight for each category is the number of speakers who have corresponding behavior normalized by the group size. For example, for the group in Figure 3(a) where all members diverge from each other, the weights will be: $converging = 0$, $diverging = 1$, and $mixedBehavior = 0$. For the group in Figure 3(b), Speaker 3 has a diverging behaviour and the rest of group have a mixed behaviour. Thus, the weights are: $converging = 0$, $diverging = 1/4$, and $mixedBehavior = 3/4$. So, the group convergence for this example is as follows:

$$Conv_{GDW} = 0 * 0 + \frac{1}{4} * C(2) + \frac{3}{4} * [C(1) + C(3) + C(4)]$$

where $C(i)$ is shortened for sum of pair convergences for speaker i normalized by the number of pairs. For example $C(2)$ is equal to:

$$C(2) = \frac{C(2,1) + C(2,3) + C(2,4)}{12}$$

5.4 EXPERIMENTS AND DISCUSSION

I break each game into four equal intervals[4] (including silences) and choose the first and last intervals to compute convergence for eight acoustic-prosodic features: maximum (max), mean, and standard deviation (SD) of pitch; max, mean, and SD of intensity; local jitter; and local shimmer. The features are extracted from each of the first and last intervals for each speaker in each team.

I evaluate the measures in two tasks that have been used for convergence measure evaluations in previous studies (De Looze et al., 2014; Lee et al., 2011; Jain et al., 2012b; Doyle et al., 2016b; Lee et al., 2011).

[4]Any method of breaking the games to compare two disjoint intervals can be used.

Predicting Team Outcomes: The first task examines how the NW, PW, and GDW measures of acoustic-prosodic convergence (independent variables) relate to the team outcome measures (dependent variables) from Section 3.1.8. This is similar to prior studies which have evaluated convergence in terms of predicting outcomes (Doyle et al., 2016b; Lee et al., 2011). I hypothesize that the group-dynamic weighted convergence measure will outperform the non-weighted and participation-based measures.

First, I train a hierarchical multiple regression with each of the three groups of convergence measures, added once in the first level and the other time in the second, to measure if the second level predictors significantly improve the explanation of variance. I only keep predictors with significant coefficients when presenting the models.[5]

For **Process Conflict,** the results show that all NW, PW, and GDW predictor groups are as good as each other; no matter which group is entered in the first level, the predictors in the second level do not significantly improve model fit.

For **Favorable**, neither PW nor NW in the second level significantly improves performance. However, Table 8 shows that adding the GDW measures at the second level significantly improves a model with only NW features at the first level. The amount of variance explained in Model 2 is significantly above and beyond Model 1, $\Delta R^2 = 0.048$, $\Delta F(2, 119) = 3.179$, $p = 0.045$. The reverse order, GDW at first level and NW at the second level, shows that the improvement at the second level is not significant, $\Delta R^2 = 0.031$, $\Delta F(2, 119) = 2.068$, $p = 0.131$. These results indicate that the proposed weighted (GDW) convergence (for intensity max and SD) are the best predictors of the favorable social outcome compared with the other two measures of convergence.

Next, I reduce the task from regression to a binary classification by splitting the two social outcome variables at the median. I perform Leave-One-Out Cross-Validations (LOOCV) using a logistic regression (L2) algorithm and all eight acoustic-prosodic features to predict binary outcomes. I use both game 1 and game 2 to increase the size of the training set (123). The results in Table 9 show that the GWD model significantly[6] outperforms both PW and

[5]To control for the effect of first versus second dialogue (game) for each group, I also included an independent variable for game. However, the coefficient was never significant.

[6]Corrected paired t-test was performed to address instance dependency from both games (Nadeau and Bengio, 2000).

	Independent Vars	M1 (β)	M2 (β)
	Intensity_max (NW)	0.248*	-0.164
	Intensity_SD(NW)	-0.055	-0.479+
	Intensity_max(GDW)		0.430+
	Intensity_SD(GDW)		0.457+
R^2		0.063	0.110
F		4.034*	3.678*

Table 8: Hierarchical regression results with intensity max and SD convergence as independent, and **Favorable** as dependent, variables. The NW measures are added in the first level and GDW measures in the second level. Significant / trending results if p-value is < 0.05 (*) or < 0.1 (+).

	Favorable	Process Conflict
Majority	50	53
NW	50	66.93
PW	53.23	**67.74+(GDW)**
GDW	**62.90****	62.90
GDW+PW	58.87	66.13

Table 9: LOOCV prediction accuracies of binary favorable social outcome and process conflict variables. (**) indicates GWD model significantly outperforms both PW and NW models. (+) indicates PW improvement over GDW is trending.

NW models to predict the favorable social outcome. In the prediction of process conflict, the PW model outperforms both NW and GDW models and its improvement over GDW is trending.

In sum, the results in both tables support my hypothesis for the favorable social outcome, where the proposed GDW convergence measure is a better predictor of the outcome. For process conflict, we do not see any significant difference.

Predicting Real Dialogues: The existence of entrainment should not be incidental. To evaluate this criteria, I use permuted versus real conversations as in (De Looze et al., 2014; Lee et al., 2011; Jain et al., 2012b). I hypothesize that GDW will be the best convergence

	All	**Game1**	**Game2**
Majority	50	50	50
NW	54.43	60.48	49.19
PW	53.62	58.06	51.61
GDW	54.03	**67.74*+**	48.39

Table 10: Accuracies using the linear SVM models and LOOCV to predict real conversations. (+) indicates GWD outperforms NW with p = 0.06 , (*) indicates GWD outperforms PW with p = 0.004.

measure for distinguishing real versus permuted dialogues.

For each of the 124 game sessions, I construct artificially permuted versions of the real dialogues as follows. For each speaker, I randomly permute the silence and speech intervals extracted by Praat. Next, I measure convergence for all the groups with permuted audios. I perform a leave-one-out cross-validation experiment to predict real conversations using the convergence measures. I examined several classification algorithms including logistic regression; linear SVM was the only one that showed significant results.

The "All" results in Table 10 show that none of the models significantly outperform the majority baseline. To diagnose the issue, I perform the prediction on each game separately. The proposed GDW model significantly outperforms other models for Game 1. However, for Game 2, none of the results are significantly different. One reason might be that convergence occurs quickly during Game 1, and there is not much convergence occurring at Game 2. Thus, there is no significant difference between permuted and not permuted convergence for any of the features during Game 2.

5.5 CHAPTER SUMMARY

In this chapter, I introduced a new weighted convergence measure for multi-party entrainment which utilizes group convergence dynamics to weight pair convergences. Experimental

results show that the proposed weighted measure is more predictive for two evaluation tasks used in prior entrainment studies: predicting favorable social outcomes and predicting real versus permuted conversations. In general, a proper weighting of the dyad-level measures performs better than non-weighted averaging in multiple tasks.

6.0 MULTI-PARTY ENTRAINMENT: FROM DIRECTIONAL MEASURES TO VECTOR REPRESENTATIONS

In previous chapters, I studied multi-party entrainment using symmetric measures such as convergance and proximity. Another category of entrainment measures are asymmetric measures (Lee et al., 2011; Danescu-Niculescu-Mizil et al., 2011, 2012; Doyle and Frank, 2016; Doyle et al., 2016a) where entrainment of A to B is not the same as entrainment of B to A. This directionality contains knowledge about the dynamics of entrainment relations in mutli-party groups and can be utilized to improve multi-party entrainment measurement. In this chapter, I utilize an existing asymmetric measure introduced in (Danescu-Niculescu-Mizil et al., 2011) to measure pairwise entrainment and propose new multi-party measures that better incorporate group dynamics.

In the first part of this chapter, I investigate whether there is a time-lag between the source and target utterances in groups using an asymmetric measure and a proposed window algorithm. In the second part, I propose a graph-based vector representation for multi-party entrainment. I incorporate the dynamics of entrainment relations estimated by the pairwise asymmetric measure.

6.1 MULTI-PARTY ADAPTIVE ENTRAINMENT AND TIME-LAG

One of the entrainment measures that has been studied for both multi-party groups and pairs is adaptive entrainment (Danescu-Niculescu-Mizil et al., 2011; Doyle and Frank, 2016). Unlike other similarity-based measures, adaptive entrainment does not simply estimate similarity in the language of the speakers but how one adapts her language to the other. The

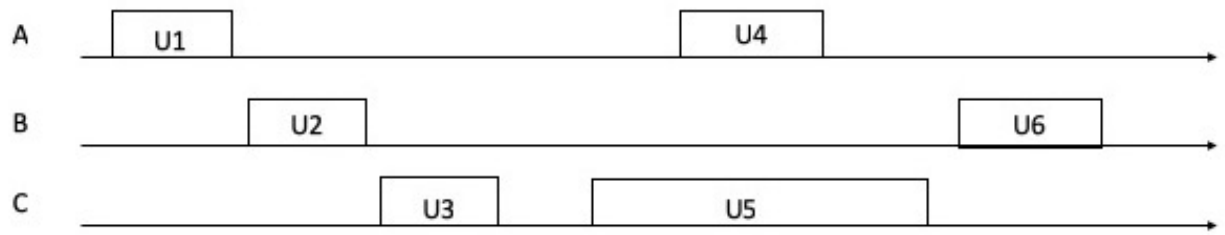


Figure 5: An example multi-party conversation. The lines show the conversation by three speakers A, B, and C. Each rectangle is an IPU.

adaptive measure that I utilize in this chapter (Danescu-Niculescu-Mizil et al., 2012) is explained in details in section 6.1.1. In adaptive measures, the similarity is the direct influence of the interlocutor's language. So, the first question, regardless of the method of calculating influence or similarity, is what is the source of influence?

Previous studies (Danescu-Niculescu-Mizil et al., 2011, 2012; Doyle and Frank, 2016) have considered the immediately preceding inter-pausal unit (IPU) (for spoken conversations) or the immediately preceding utterance (for the text-based conversations) of the interlocutor as the source of influence for the current IPU/utterance. But, a spontaneous face-to-face multi-party spoken conversation is more complicated.

First, detecting the immediately preceding IPU is not always straight-forward. Consider the example in Figure 5. Each block represents an IPU. The immediately preceding IPU of U_3 is U_2. But, how do we choose the immediately preceding IPU of U_6? Do we choose the IPU that ends last or starts last?

Second, in a two-party conversation, the addressee is always the other speaker. But, in a multi-party conversation, the current utterance can be the response to any of the other speakers (not necessarily the speaker with the immediately preceding IPU), to the group, or a subgroup. Similarly, the source of influence for the target IPU might be other than the immediately preceding IPU. In other words, there might be a time-lag between the source IPU and the target IPU in a multi-party spoken conversation.

So, from this perspective, choosing the immediately preceding IPU as the only source of entrainment for the current IPU might be too restrictive. Moreover, considering all preceding utterances as the potential source of entrainment is too simplistic. Addressee detection for multi-party human-human conversations is a complex task which works best if visual information like gaze is available (Jovanovic et al., 2006). Moreover, automatic addressee detection requires annotated data. Finally, there is no evidence that other conversational roles such as side-listeners do not entrain to the speaker. Recency (Brennan and Clark, 1996) might be a more important factor than conversational roles. So, I employ a temporal window approach to consider recent IPUs as the potential source of entrainment for each target IPU. I set-up experiments to examine the effect of the length of the window or the amount of time-lag on multi-party entrainment.

In the next section, I describe the adaptive entrainment measure (Danescu-Niculescu-Mizil et al., 2012) that I utilized in this chapter. In section 6.1.2, I describe the proposed window-based approach to build the source-target pairwise sets.

6.1.1 Probabilistic Adaptive Entrainment

The measure, introduced by Danescu-Niculescu-Mizil et al. (2012), is based on conditional probabilities. The entrainment of speaker b to speaker a with respect to a linguistic feature c, $Ent_c(b, a)$, is defined in equation 6.1:

$$Ent_c(b, a) = p(e_b|e_a) - p(e_b) \tag{6.1}$$

where the probabilities are estimated over the set U_{ab}. U_{ab} is the set of all (u_a, u_b) conversation exchanges where speaker a's utterance (u_a) is uttered immediately preceding to the speaker b's utterance (u_b). I call u_a source and u_b target. e_b/e_a is the indicator that the desired event (*e.g.* presence of linguistic feature c) occurred in the corresponding utterances u_b /u_a from set U_{ab}. "This measure estimates how much a's use of c in an utterance u_a increases the probability that b will use c in his reply u_b, where the increase is relative to b's normal usage of c in conversations with a. "

Similarly, entrainment of speaker b towards a group of speakers G in a multi-party

conversation is defined in equation 6.2:

$$Ent_c(b, G) = p(e_b|e_G) - p(e_b) \tag{6.2}$$

where U_{Gb} is the set of (u_G, u_b) exchanges which involves first utterances u_G from various speakers in G. Finally, the multi-party entrainment is the average of the entrainment of all speakers to the group.

$$Ent_c(G) = \left\langle Ent_c(b, G) \right\rangle b \in G \tag{6.3}$$

In the next section, I describe how I define set U_{Gb} using the proposed window-based approach to consider the possible time-lag of influence.

6.1.2 Proposed Window-Based Approach

To measure entrainment, we need to build the set of source and target IPUs U. Instead of only considering the immediately preceding IPU, I propose to consider a window of length L. The proposed method is defined in Algorithm 1. First, I sort all IPUs by their start time. Then, for each target IPU, I consider a window of length L that starts and ends at $[Target.start - L, Target.Start)$. All IPU's in this window which are not uttered by the target IPU's speaker can be a potential source. So, for each pairwise set U_{ab}, I add an exchange (u_a, u_b) to the set where the first element includes concatenation of all IPU's uttered by a in the window. In this way, each event is only considered once in a window for a pair. For each group-wise set U_{Gb}, I add an exchange (u_G, u_b) to the set where the first element includes concatenation of all IPU's not uttered by b in the window. The size of the window is the parameter that indicates the time-lag of influence and needs to be tuned. This is an approximate algorithm. I make sure that the source IPU's start time is smaller than the target IPU's start time. But, I don't force them to end before the target IPU's begin. The reason is that there is a lot of overlapping speech which we do not want to discard. For this purpose, I only make sure that at least half of the length of each IPU is uttered before the target's start time.

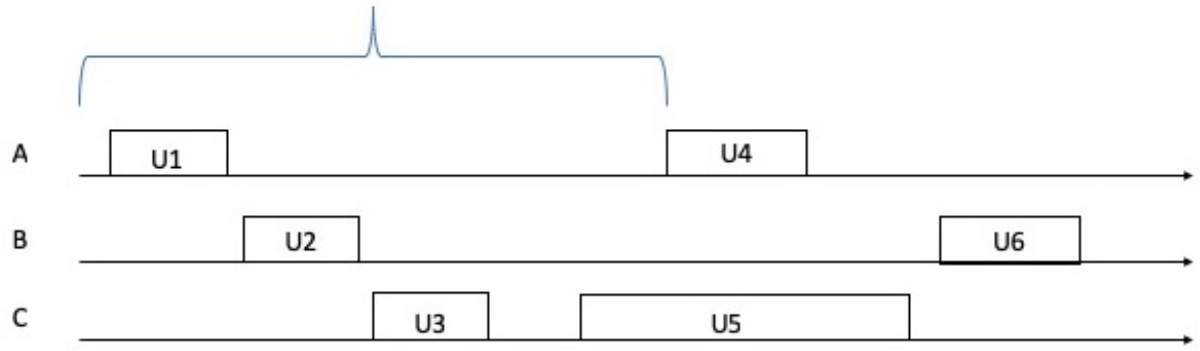


Figure 6: An example multi-party conversation. The lines show the conversation by three speakers A, B, and C. Each rectangle is an IPU. The bracket shows a window.

Consider the example in Figure 6. Suppose that the current target IPU is U_4 and the bracket shows the corresponding window for U_4 with length L. The source-target pairs where target is U_4 are:

- U(B,A) += (U_2, U_4)
- U(C,A) += (U_3, U_4)
- U(group,A) += $(U_2 + U_3, U_4)$

U_1 is uttered by A and is ignored. U_5 does not satisfy the condition in line 8 of the algorithm and is ignored. Considering entrainment of A to the group, I concatenate all valid utterances from all speakers (U_2+U_3). + = indicates that the final U(B,A) set is the union of all relevant source-target pairs in the conversation.

6.1.3 Experiments and Discussion

I compare the proposed window-based method and the baseline immediately preceding IPU using the probabilistic measure defined in Equation 6.2. For the baseline method, I choose the immediately preceding IPU with the biggest start time, unless it does not pass the condition in line 8 of the window algorithm. Also, I add all the other IPUs overlapping with the

Algorithm 1 Window-Based Pair Construction

1: *Sort the IPUs by their start time*
2: $L \leftarrow$ window length
3: **for all** targetIPU $\in$ set (sorted IPUs) **do**
4: $u, speakers \leftarrow$ Empty
5: Window $\leftarrow$ all IPUs that their start time is in $[targetIPU.start - L, targetIPU.start)$
6: **for all** sourceIPU $\in$ window **do**
7: **if** sourceIPU.speaker ! = targetIPU.speaker **then**
8: **if** sourceIPU.start + sourceIPU.length / 2 < targetIPU.start **then**
9: u(sourceIPU.speaker) += sourceIPU
10: u(group) += sourceIPU
11: speakers.append(sourceIPU.speaker)
12: **for all** s $\in$ speakers + ['group'] **do**
13: U(s,targetIPU.speaker).append((u(s),targetIPU))
14: **return** U

selected preceding IPU if their end time is bigger than the selected IPU's end time and if they pass the condition in line 8 of the algorithm. So, I include the IPUs if they start or end last. I experiment with different values of window length to test the hypothesis that there is a time lag between the source and target of adaptive entrainment. For consistency with prior work (Danescu-Niculescu-Mizil et al., 2012), I measure lexical entrainment on eight LIWC-derived categories of function words (Pennebaker et al., 2001) that have little semantic meaning and is more relevant to style than content. These eight categories are: articles, auxiliary verbs, conjunctions, high-frequency adverbs, impersonal pronouns, personal pronouns, prepositions, and quantifiers (451 lexemes total). I use the multi-party entrainment on these eight categories as features to predict team outcomes (process conflict and favorable outcome). The two tasks are binary classifications. I use Leave-One-Out cross validation and support vector machine with the RBF kernel.

The results in Table 11 are the accuracies. I examine the window length as low as four seconds and as high as 60 seconds. The average silence length in the dataset is 1.11

Features	Process Conflict	Favorable Outcome
Majority	53.78	50.42
Baseline	63.02	51.26
Window (L=4s)	55.46	47.89
Window (L=5s)	59.66	53.78
Window (L=6s)	53.78	50.42
Window (L=7s)	63.02	47.89
Window (L=8s)	57.98	59.50
Window (L=9s)	57.98	48.73
Window (L=10s)	57.98	48.73
Window (L=15s)	63.86	46.21
Window (L=20s)	62.18	46.21
Window (L=25s)	64.70	39.49
Window (L=30s)	65.54	45.37
Window (L=40s)	64.70	50.42
Window (L=50s)	63.02	49.58
Window (L=60s)	63.86	48.73

Table 11: Accuracy of binary classification of process conflict and favorable outcome for temporal window approach.

seconds and the maximum silence length is 31.97 seconds. I choose the minimum and maximum window length with regards to these numbers to avoid empty windows. There is no significant difference between the results. On predicting conflict, larger window sizes (> 10) performs better than small window sizes. The window length of 30 seconds has the best performance although the difference with baseline is not significant. We do not observe the same pattern on predicting favorable outcome. On this task smaller window sizes (< 9) performs better. Looking at these results, we cannot make a consistent conclusion for both tasks.

Features	Is Real?
Majority	50.00
Baseline	83.61
Window (L=4s)	75.63
Window (L=5s)	79.41
Window (L=6s)	77.73
Window (L=7s)	78.15
Window (L=8s)	78.15
Window (L=9s)	77.73
Window (L=10s)	73.10
Window (L=15s)	75.63
Window (L=20s)	73.95
Window (L=25s)	73.10
Window (L=30s)	66.38
Window (L=40s)	64.70
Window (L=50s)	65.12
Window (L=60s)	61.70

Table 12: Accuracy of binary classification of real or permuted games for temporal window approach.

Table 12 shows the results of validation task (real vs permuted prediction). The results indicate that increasing the window size decreases the accuracy. By increasing the window size, we are giving more opportunity to speaker's to influence their speaking partners. In the Equation 6.2, increasing the window size only influences the first term and potentially increases the probability. So, why does the quality of the real vs permuted task decrease? Increasing the window size, entrainment of the permuted games are potentially increased and as a result the entrainment values are less distinctive from the entrainment values of real games.

6.1.4 Summary

I hypothesized that in adaptive entrainment, influence might occur within a time lag and finding the effective time-lag will increase the quality of multi-party entrainment and as a result the quality of predicting social outcomes using entrainment. To test this hypothesis, I proposed a temporal window-based approach to build the source and target pairwise sets. I used these sets to measure entrainment by Equation 6.2 and used simple averaging to measure multi-party entrainment. The results of the three prediction tasks (process conflict, favorable outcome and fake vs real) do not support my hypothesis that the time-lagged measure outperforms the baseline. There is no significant improvement as a result of increasing time-lag. There are some non-significant improvements but the results are not consistent across all these tasks. So, in the next section, I use the immediately preceding IPUs baseline (with the biggest start time and their overlapping IPUs) as explained in the previous section.

6.2 VECTOR REPRESENTATION FOR MULTI-PARTY ENTRAINMENT

In this section, utilizing the adaptive pairwise entrainment measure (Danescu-Niculescu-Mizil et al., 2012), which measures how much a speaker adapts her language to another one in a local turn-by-turn basis, I propose a graph-based vector representation of multi-party entrainment to encode the strength and structure of pairwise interactions in multi-party groups. Weighted directed entrainment graphs represent the structure of pairwise entrainment relations and their strength. Learning embedding for the entrainment graphs, I represent multi-party entrainment in vector-space where groups with similar graphs have close vectors.

Learning dense vector representations for nodes, edges, or sub-graphs from a large-scale sparse graph has been studied by researchers and applied to social or knowledge graphs (Luo et al., 2015; Tang et al., 2015; Grover and Leskovec, 2016; Zhou et al., 2017; Hamilton et al., 2017). The intuition is similar to *word2vec* (Mikolov et al., 2013). Similar nodes in a graph should have vector representations that are close to each other. Similarity can be defined as

having similar neighbors or similar structural roles.

Inspired by these methods and by *paragraph2vec* (Le and Mikolov, 2014), I learn vectors for small graphs where similarity is defined as having a similar graph structure. To encode the structure, I propose to initialize the node and graph embeddings by applying a set of graph algorithms where each encodes a distinctive property of the graph. As the supervision, I propose to employ the domain-specific task of predicting real versus randomly permuted conversations, which has been utilized in the entrainment domain to verify the validity of measures (De Looze et al., 2014; Lee et al., 2011; Jain et al., 2012b; Rahimi and Litman, 2018). Experimental evaluations demonstrate that the group entrainment embedding improves performance for the downstream task of predicting group outcomes compared to the state-of-the-art methods.

In the next section, I describe entrainment graphs. Then, in the following section, I describe the three proposed approaches to represent entrainment in a vector space utilizing the graphs.

6.2.1 Entrainment Graph

Influence networks or graphs have been introduced and investigated in other domains such as the social analysis literature (Friedkin and Johnsen, 2011; Tang et al., 2009; Romero et al., 2011). I propose to apply a similar idea to build multi-party entrainment graphs. First, I estimate pairwise entrainment values using an existing probabilistic directional (i.e., asymmetric) method (Danescu-Niculescu-Mizil et al., 2012), where the entrainment of speaker b to speaker a on a lexical category or linguistic feature c, $Ent_c(b, a)$, is defined as in Equation 6.1. The directionality of this entrainment measure means that the entrainment of speaker a towards b is not the same as the entrainment of speaker b towards a.

Next, I define the multi-party entrainment graph $G = (V, E, W)$. Each node $v \in V$ represents a speaker from the group. The directed edges in E represent the presence and the weights in W represent the strength of entrainment between the source and destination node, which are measured using Equation 6.1. I define an edge from node a to b if and only if $Ent_c(b, a)$ is positive. The negative values imply that the linguistic feature c is part of the

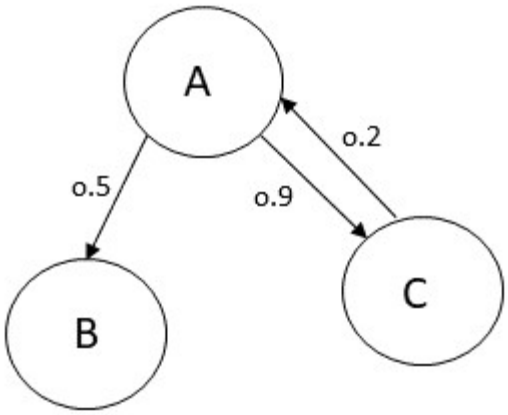


Figure 7: An example multi-party entrainment graph

speaker's b language and is independent of speaker a's speech. So, I can simply ignore them. The direction of the edge implies that the source node influences the destination node, while the edge weight represents the amount of this influence (entrainment). Consider a group with three speakers: A, B, C. Suppose that the entrainment of pairs on a lexical feature are as follows:

$$Ent(B, A) = 0.5,$$
$$Ent(A, B) = -0.1,$$
$$Ent(C, A) = 0.9,$$
$$Ent(A, C) = 0.2,$$
$$Ent(B, C) = -0.5,$$
$$Ent(C, B) = 0.$$

The entrainment graph that would then be constructed for this group is shown in Figure 7.

Entrainment graphs contain interesting information about the dynamics of the entrainment relations. For example, we could learn the structural roles of the speakers, such as

who are the influencers, connectors, or passive speakers. Also, we could learn about indirect entrainment relations. If A influences B and B influences C, there is an indirect influence from A to C. This information could potentially help us have a better understanding of multi-party entrainment. In the graph from Figure 7, we observe that A is an influential speaker and C could potentially influence B although there is no edge between them. This information will be lost if we simply average pairs' entrainment values.

In the next section, I propose three different approaches to learn entrainment vector representations from such entrainment graphs.

6.2.2 From Graphs to Vector Representation

6.2.2.1 Directly Estimating the Vectors

Given the entrainment graphs, the goal is to represent group entrainment in a vector-space where groups with similar entrainment dynamics have close vectors. As the most obvious and straightforward approach, I apply a set of graph algorithms to capture distinctive and informative properties of the graphs. Applying d functions, each node in the graph (i.e., conversational participant) is represented with a d-dimensional vector. Then, the vector representation of the entrainment graph (i.e., the group of participants) is a simple average of its nodes' vectors.

Following are all ten kernel functions that I utilized. I tried to be inclusive but I did not perform an experiment to obtain the best list by pruning it or by adding other algorithms. For convenience, I call these functions kernels in the rest of this thesis. All these functions are well-known graph algorithms, so I explain them minimally here. Given a weighted directed graph $G = (V, E, W)$, the 10 kernels are as follows:

- **K1= Closeness Centrality** (Wasserman and Faust, 1994) of a node u is "the ratio of the fraction of reachable nodes, to the reciprocal average distance from the reachable node".

$$C(u) = \frac{n-1}{N-1} \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \tag{6.4}$$

The distance function $d(v, u)$ is defined as the length of the shortest-path from node v to u. And, n is the number of nodes that can reach u. N is the total number of nodes. In this work, we are interested in finding the highly influential nodes which can reach to many nodes. This is opposite of the Closeness Centrality definition. For this purpose, for each entrainment graph G, I build and use the reversed graph G^R in which the edges are the reverse of G's edges. Thus, the reachable nodes of u in G^R are the ones that can be reached starting from u in the original graph G. So, highly influential nodes in G have high closeness centrality score.

- **K2 = Betweenness Centrality** (Brandes, 2001) of node u is the fraction of all-pairs shortest paths that pass through u to all the all-pairs shortest paths. $\sigma(s, t)$ is the number of shortest (s, t)-paths, and $\sigma(s, t|u)$ is the number of those paths passing through node u other than s, t. If $s = t$, $\sigma(s, t) = 1$, and if $u \in \{s, t\}, \sigma(s, t|u) = 0$.

$$C_B(u) = \sum_{s,t \in V} \frac{\sigma(s, t|u)}{\sigma(s, t)} \tag{6.5}$$

  A node has a high betweenness centrality if it's role in the graph is a connector.

- **K3 =PageRank** [1] (Page et al., 1999) is one of the most well-known algorithms which was originally designed to rank web pages. It outputs a probability distribution for nodes based on the score of their neighbours. "PageRank works by counting the number and quality of links to a node to determine a rough estimate of how important the node is. The underlying assumption is that more important nodes are likely to receive more links from other nodes." I use the weighted reversed graph G^R, and the weighted algorithm (Xing and Ghorbani, 2004) to measure PageRank probability distribution of the nodes.

- **K4 , K5 = HITS** (Kleinberg, 1999) computes two numbers for a node. Authority score is based on the incoming links ("a good authority is a node that is linked by many different hubs"). Hub score is based on outgoing links ("a good hub is a node that points to many other nodes").

[1] PageRank and HITS algorithms' outputs are a probability distribution. So, although they are distinct and informative at the node level, they are repetitive and all equal to the team-size at the graph level. For the purpose of consistency with the next proposed method, I include them all here.

- **K6 = Maximum Flow** (Ford and Fulkerson, 2009) of node u is the sum of all single-commodity max flow values (i.e., net outflow) from the source node u to all other nodes in the capacity graph $G^c = (V^c, E, C)$ where each edge has only a single attribute, capacity C, which is equal to the weight attribute in the original graph G.

$$MaxFlow(u) = \sum_{v \in V^c} max\_flow(u, v, C) \tag{6.6}$$

  The Maximum Flow, is the sum of all direct and indirect entrainment influences that a node has on all other nodes in the graph.

- **K7 = Katz Centrality** (Katz, 1953) computes centrality for a node based on centrality of its neighbors. The Katz centrality for node u is:

$$KATZ_u = \alpha \sum_j A_{ij} KATZ_j + \beta \tag{6.7}$$

  where A is the adjacency matrix of graph G with eigenvalues λ. The parameter β controls the initial centrality and $\alpha < \frac{1}{\lambda_{max}}$. "Katz centrality computes the relative influence of a node within a network by measuring the number of the immediate neighbors (first degree nodes) and also all other nodes in the network that connect to the node under consideration through these immediate neighbors."

- **K8 = Weighted In_degree** of a node is sum of the weights of all incoming edges of it. This indicates how much direct influence the node gets from other nodes in the graph.

- **K9 = In_degree Centrality** of a node is the fraction of nodes to which it's incoming edges are connected. This measure indicates how many influencers a node has.

- **K10 = Degree Centrality** of a node is the fraction of nodes that it is connected to (sum of in_degree and out_degree centrality). This measure indicates the ratio of all the nodes that entrain directly to/from the corresponding node.

6.2.2.2 Learning Embedding: The Self-Supervised Approach

Given the entrainment graphs, the goal is to learn a d-dimensional embedding of nodes and graphs where nodes with similar structural roles and graphs with similar structure have

close vectors. I employ the *node2vec* (Grover and Leskovec, 2016) and *paragraph2vec* (Le and Mikolov, 2014) methods and define our problem as follows.

Our embedding learning is a maximum likelihood optimization. I define G as the set of all graphs and U as the set of all nodes (vocabulary) from all graphs. Then, for a given graph $g = (V, E, W)$, the nodes of the graph, V, are its context, $C(g) = V$. I seek to optimize the objective function in Equation 6.8 which maximizes the log-probability of observing the context of the graph g, conditioned on vector representation of g, given by $f(g)$:

$$max_f \sum_{g \in G} \log p(C(g)|f(g)) \tag{6.8}$$

Assuming conditional independence of observing each node in the context given the vector representation of the graph, the conditional probability of Equation 6.8 is defined by:

$$P(C(g)|f(g)) = \prod_{v \in C(g)} p(v|f(g)) \tag{6.9}$$

Let W be the matrix of graph embedding and Z be the matrix of node embedding. Every unique graph g is mapped to a unique vector W_g and every unique node is mapped to a unique vector Z_v. The probability in Equation 6.9 is defined as the softmax probability normalized by all the nodes in U:

$$p(v|f(g)) = \frac{\exp(Z_v^T.W_g)}{\sum_{u \in U} \exp(Z_u^T.W_g)} \tag{6.10}$$

Given Equations 6.8, 6.9, and 6.10, and approximating the normalization of the softmax, which is a sum over all nodes of all graphs, with negative sampling (Mikolov et al., 2013; Grover and Leskovec, 2016). the loss function of the optimization problem is:

$$L = -\sum_{g \in G} \sum_{v \in C(g)} (\log \sigma(Z_v^T.W_g) + \sum_{u \in C'(g)} \log \sigma(Z_u^T.W_g)) \tag{6.11}$$

The dot product measures the similarity of a node and a graph in the vector-space. For all nodes in the context of a graph, the vector representations of the nodes and the graph should be close in the vector space. $C'(g)$ is a set of random nodes which do not belong to the graph g. σ is the sigmoid function.

Initializing Embedding: As discussed before, we want graphs with similar structure and nodes with similar structural roles to be close in vector space. To achieve this, we need to encode these structures in the vectors. For this purpose, I utilize the proposed kernel approach from Section 6.2.2.1 to initialize the node and graph embedding matrices. So, the vector of each node or graph indicates their structural properties and similarity of vectors indicates similarity of their structures.

Domain-Specific Negative Sampling: To build the set $C'(g)$ in Equation 6.11, one approach similar to *word2vec* is to randomly sample nodes that are not in the context of the graph (i.e., that are from other graphs). But, this might not be a good approach for our problem as several graphs might have the same structure. So, randomly choosing nodes from other graphs does not serve our purpose well. We want our negative samples to have different vector representations from the context of the graph.

Entrainment is a phenomena that occurs in the course of conversation. So, randomly generated conversations should not show strong entrainment relations. Distinguishing between real and randomly permuted fake conversations is a validation task in the entrainment literature (De Looze et al., 2014; Lee et al., 2011; Jain et al., 2012b; Rahimi et al., 2017). I thus propose to use the permuted version of each conversation to build the corresponding fake graphs and use the nodes of these fake graphs to build $C'(g)$. So, the size of $C'(g)$ is equal to $C(g)$ and the nodes in $C'(g)$ are the permuted version of the nodes in $C(g)$. The fake conversations were generated by randomly permuting the speech and silence intervals of each speaker in the group. Using this method, I make sure that the negative and positive samples have different structures. At the same time, I make sure they are not too distinct since I do not change the content of the conversations but only randomly permute the content. For example, the distribution of the lexical categories are the same in both negative and positive samples of a graph.

Network Structure and Algorithm: The network is in Figure 8. As in *word2vec* (Mikolov et al., 2013), our network is a shallow two-layer neural net. It has two embedding look-up tables, one for the nodes and one for the graphs. The second (output) layer includes a sigmoid activation function applied on the dot product of the two vectors from the first layer. The input is a pair of node index and graph index. Given the input indexes and

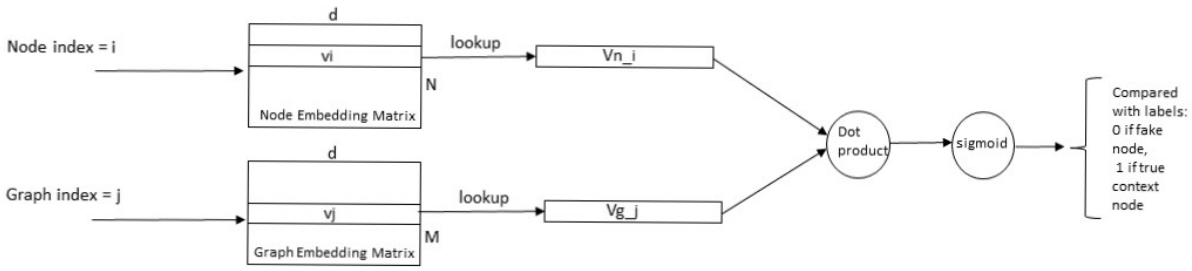


Figure 8: GraphVec

the embedding matrices, I look up the two vectors for the given node and graph. Then, I simply get the dot product of the two vectors to calculate their similarity. I apply a sigmoid activation function on the result of the dot product to calculate the probability of the output. This probability is used to predict if the node is from the context of this graph (a positive sample) or is from the corresponding fake graph (a negative sample). So, I have a simple binary prediction and I use binary cross entropy loss function and a stochastic gradient decent optimization algorithm. After completing the training, I get the learned graph embedding matrix as our representation of multi-party entrainment.

The pseudo-code of the algorithm is given in Algorithm 2.

6.2.2.3 Learning Embedding: The Weakly-Supervised Approach

The self-supervised approach employs the randomly permuted conversations to generate the negative samples for an optimization task that tries to maximize the likelihood of observing the context nodes. A different approach is to directly employ the validation task of real-vs-permuted prediction as supervision. Given vector representations of a node and a graph, I predict if the input is from a real conversation or a fake conversation. I train the graph embedding while I optimize this classifier. I call this method weakly supervised since the supervision is not on the main task of interest which is predicting social outcomes and

Algorithm 2 Learning Multi-party Entrainment Embedding

```
1: function InitializeEmbedding(Realgroups , PermutedGroups)
2:     for all group ∈ Realgroups + PermutedGroups do
3:         pairEnts ← PairEnrtainments(group)
4:         graph ← BuildGraphs(pairEnts)
5:         vector ← ApplyKernel(graph)
6:         NodeEmbedding ← append(NodeEmbedding,vector)
7:     GraphEmbedding ← GetGraphEmbeding(NodeEmbeding)
8:     return NodeEmbedding, GraphEmbedding
9: function LearnEmbedding(input,labels,NodeEmbedding,GraphEmbedding)
10:    Initialize Embedings to NodeEmbedding,GraphEmbedding
11:    model ← StochasticGradiantDesent(input,label)
12:    LearnedGraphEmbeding ← model.GetWeights(Graph Embedding Layer)
13:    return LearnedGraphEmbeding
```

also no manually labeled data is required. Similar to the self-supervised approach, I construct a randomly permuted conversation for each real conversation and build its corresponding entrainment graph. Then, unlike the self-supervised approach, I predict if a given input graph is real or permuted. The objective is to minimize a binary cross entropy loss function or maximize the likelihood of observing the data using a gradient descent optimization.

Like before, I employ two embedding matrices for the nodes and the graphs although the size of the graph embedding is doubled (by including the fake graphs). I initialize the two matrices with the kernel approach to encode the structure of the graphs. Assuming that the kernel initialization is good at encoding the underlying structure, to reduce the number of trainable parameters, I fix the node embedding with the initialization, and train the graph embedding. I utilize the same shallow network as in the self-supervised method. After training the graph embedding, the subset of the graph embedding which is from the real graphs represents our entrainment vectors.

6.2.3 Evaluation of Multi-party Entrainment Embedding

6.2.3.1 Experimental Setups

For consistency with prior work (Danescu-Niculescu-Mizil et al., 2012; Doyle and Frank, 2016), I measure lexical entrainment on eight LIWC-derived categories of function words (Pennebaker et al., 2001) that have little semantic meaning and are more relevant to style than content. These eight categories are: articles, auxiliary verbs, conjunctions, high-frequency adverbs, impersonal pronouns, personal pronouns, prepositions, and quantifiers (451 lexemes total). Transcripts are pre-processed before extracting lexical terms by removing punctuation marks, converting all words to lower case, removing noises such as laughter, and removing any part of the transcript indicated as not fully understood by transcribers.

The size of the input data for training the networks is about 6500 instances[2]. I have minimum 100 and maximum 150 epochs and stop early if the training loss is smaller than 0.1 to avoid overfitting. The batch size is set to 20. *RMSProp* optimization algorithm is used with learning rate equal to 0.001. These hyper-parameters are chosen with regards to the small training data size and should be optimized in the future. It should be emphasized that I do not require any manually labeled data for training the self-supervised or weakly-supervised models.

After learning the entrainment embedding, similar to prior works (De Looze et al., 2014; Lee et al., 2011; Jain et al., 2012b; Rahimi et al., 2017; Doyle et al., 2016b; Lee et al., 2011) which have evaluated entrainment in terms of predicting outcomes, I evaluate its utility at predicting *Favorable* and *Conflict* team outcomes. I use support vector machines with RBF kernel and perform leave-one-out cross validation. The size of the data for this experiment is 119 since I predict the outcomes for each team and each session. I compare the utility of the proposed embedding with two local adaptive baselines from the literature: SCP (Danescu-Niculescu-Mizil et al., 2012) which is a probabilistic measure and its pairwise version is utilized in this thesis to build the entrainment graphs, and HAM (Doyle and Frank, 2016) which is a generative hierarchical alignment model argued to outperform SCP. The HAM baseline is fit using the same hyper-parameters as in (Doyle and Frank, 2016). But, it is fit

[2] 119 sessions * (3 or 4) speakers * 8 LIWC categories * 2 (fake or real)

with 2000 iterations of the sampler (1000 as warm-ups) and four chains since our data is small. The output is a probability distribution for each group. So, I utilize the mean, upper and lower bounds of the 95% highest posterior density.

6.2.3.2 Results and Discussion

First, I utilize the entrainment embedding of all 8 lexical categories as features. So the total number of features is 80 which for 119 instances of data is a lot and might cause the model to overfit. So, it is important to employ feature selection. I employ a model-based feature selection using LASSO. The feature selection is performed inside the cross validation loop, so the number of selected features might be slightly different at each fold. I also set a threshold on minimum number of features to avoid underfitting. The regularization parameter is tuned with regards to this threshold. The threshold itself is tuned in each fold.

The results are in Table 13. The middle part of the table shows the accuracies of the three proposed approaches using their best configurations at two tasks of predicting Favorable and Conflict outcomes using all features or by employing feature selection. Predicting Conflict, the embedding of weakly-supervised approach when utilizing feature selection outperforms the best SCP result significantly and the best HAM result. Feature selection has a greater effect on the proposed approaches since the number of features in these models are more than baselines. SCP has only 8 features and HAM has 24 features. Predicting Favorable outcome, the embedding of weakly-supervised approach with or without feature selection outperforms the best SCP result significantly and the best HAM result with a trending p-value (< 0.1). Comparing the three proposed approaches, the weakly-supervised approach outperforms the other two models on both tasks. This shows that training the embedding improved the initial kernel embedding although the initial kernel embedding, which does not require any training, performs comparable, if not better, to the computationally expensive HAM baseline.

The last part of Table 13 presents other evaluated configurations for the weakly-supervised and the self-supervised approach. For the self-supervised approach, I experiment on initializing the graph embedding with the default **U**niform initialization rather than our proposed

	Features	Conflict-All	Conflict-FS	Favorable-All	Favorable-FS
Baselines	Majority	53.78	53.78	50.42	50.42
	SCP	63.02	63.86	51.26	50.42
	HAM	68.90	69.74	53.78	51.26
Best Proposed	Kernel	57.14	69.74	$62.18^{(*,)}$	$62.18^{(*,)}$
	WeakS_KP_NT	59.66	$\mathbf{74.79}^{(*,)}$	$\mathbf{63.86}^{(*,+)}$	$\mathbf{63.86}^{(*,+)}$
	Self_K_P	63.02	$\mathbf{73.95}^{(+,)}$	57.14	58.82
Other Configurations	WeakS_KP_T	58.82	63.86	56.30	57.14
	Self_K_R	63.02	66.38	56.30	55.46
	Self_U_P	49.58	63.86	39.49	52.94
	Self_U_R	52.94	57.93	58.82	55.46

Table 13: Accuracy of predicting Conflict and Favorable outcomes. The features are entrainment values/vectors from all 8 LIWC categories. The pair of signs in parenthesis indicates the result of significant test comparing the corresponding accuracies with the best of SCP and HAM in order. "*" indicates significance (p-value < 0.05), "+" indicates trending result (p-value < 0.1).

Kernel initialization. Also, for negative sampling, I experiment on **R**andomly selecting nodes from other graphs rather than our proposed domain-specific **P**ermuted negative samples. The results show that both proposed approaches for initialization and negative sampling outperform the other configurations. For the weakly-supervised approach, the best configuration has Not Trainable (**NT**) node embedding and only trains the graph embedding rather than **T**raining both embedding matrices. This is beneficial since it reduces the number of trainable parameters of the model.

In a second experiment, I predict the Favorable and Conflict outcomes using the embedding of each lexical category. So, I have 8 prediction tasks for each outcome. At each prediction task, the number of features is equal to the size of the vector (10). So, I do not need to employ feature selections in this experiment. The results are in Table 14. I only experiment with the best methods from the first experiment. For predicting Conflict, I observe that the HAM baseline is more robust across different lexical categories than the embedding approaches. But, the proposed approaches outperform the HAM baseline on all categories when predicting favorable outcome. So, the entrainment embedding is more robust across the two tasks.

Given the promising results of the proposed vector representations, there are two more questions that I further investigate. First, does the proposed model learn anything beyond team size? [3] Second, which dimensions (kernels) are more predictive? To answer these questions, I take a closer look at each dimension of the vectors. I choose one of the experiments where the vector representation outperformed baselines on a individual category: kernel model predicting process conflict on the LIWC category of "quantitative". I perform a hierarchical regression analysis. The z-scored team-level process Conflict is the dependent variable. I enter team-size as an independent variable to the first level. In the second level, I add all ten dimensions of the entrainment vector on the LIWC category of *quantitative*.

The results are in Table 15. I remove all the dimensions that did not have a significant or trending coefficient from the final model. The amount of variance explained by Kernel dimensions is significantly above and beyond team size entered in Model 1, $\Delta R^2 = 0.215$,

[3] I specially need to answer this question for the Kernel approach, since average of PageRank and HITS scores over the nodes of the graph is the ratio of the number of nodes. This is not an issue for the learned embedding.

Task	Features	ipron	article	auxverb	conj	adverb	ppron	preps	quant
Conflict	Majority	53.78	53.78	53.78	53.78	53.78	53.78	53.78	53.78
	HAM	**68.06**	**69.74**	63.86	**68.06**	**61.34**	**67.22**	**64.70**	65.54
	Kernel	62.18	63.86	58.82	64.70	**61.34**	**66.38**	60.50	65.54
	WeakS_KP_NT	57.98	57.14	**65.54**	**67.23**	56.30	60.50	59.66	**72.27**
	Self_K_P	**68.06**	54.62	63.02	57.14	58.82	53.78	55.46	**71.42**
Favorable	HAM	50.42	54.62	45.37	43.69	40.33	47.05	37.81	51.26
	Kernel	**63.02**	63.02	**61.34**	58.82	**55.46**	**52.94**	**63.02**	54.62
	WeakS_KP_NT	58.82	**66.39**	57.14	57.14	**55.46**	51.26	**63.03**	**57.98**
	Self_K_P	57.14	53.78	48.73	**63.86**	54.62	51.26	47.05	49.58

Table 14: Accuracy of predicting Conflict and Favorable outcomes. The features are entrainment values/vectors from a single lexical category.

	Model 1			Model 2		
	B	SEB	β	B	SEB	β
Team Size	0.613	0.180	0.300*	0.510	0.210	0.250*
K1 = Closeness Centrality				18.922	9.201	4.509*
K2 = Betweenness Centrality				-5.942	3.563	-0.526+
K6 = Maximum Flow				-4.280	1.763	-0.268*
K10 = Degree Centrality				-8.281	4.588	-3.584+
Model R^2		0.09			0.306	
Model F		11.609*			9.948*	

Table 15: Summary of hierarchical regression analysis for variables predicting process conflict on quantitatives using Kernel approach. B , SEB, and β are the unstandardized coefficients, coefficients Std. Error, and standardized coefficients. * $p < .05$. + $p < .1$ $n = 119$.

$\Delta F(4, 113) = 8.763$, $p = 0.000$. So, the answer to our first question is yes. The proposed model learns predictive dimensions above and beyond the effect of team size. To answer the second question, I look at the selected dimensions with significant coefficients. Closeness Centrality has a positive significant correlation with process conflict which means teams with higher average closeness centrality have higher conflict. In other words, teams with fewer influential members, have less conflict. Maximum flow has a significant negative correlation which indicates teams with higher average maximum flow, have less conflict. In other words, the more is the direct and indirect entrainment in the team, the less is the process conflict. One main advantage of the proposed vector representation compared to the baselines is the ability to present all these pieces of information about the dynamics of entrainment in groups.

6.2.4 Summery

In this section, I proposed group entrainment embedding, a vector representation for multi-party entrainment to encode the underlying entrainment dynamics in the groups. I proposed three approaches to learn the vector representation from entrainment graphs built by utilizing existing directional pairwise entrainment measures. I concluded that the vector representation learned by the proposed weakly-supervised approach outperforms the baselines and the other proposed approaches. Beside performance, this approach has other advantages. First, encoding the underlying structure of the entrainment graphs in the vectors provides useful information. For example, I found that teams with more influential (in terms of entrainment) members or higher average closeness centrality have more Process Conflict. Second, proposed approaches are computationally less expensive than the best performing baseline: the generative HAM model. The running time of the weakly supervised training on a laptop was less than 30 seconds while HAM takes about 30 minutes to converge running on a server. Finally, the weakly supervised approach requires training data similar to HAM. But, the proposed kernel approach when performance is comparable to the weakly supervised approach does not require any training data and directly estimates entrainment of groups.

I did not perform any parameter tuning on the parameters of the neural network such

as number of epochs, batch size, or learning rate. I also chose a simple two layer network. Other network structures might perform better specially for the weakly-supervised approach. Further investigation is required to optimize the list of the graph algorithms (kernels) to best encode the structure of the graphs. The results are promising and might be even further improved by exploring these paths. Also, the team size in the Teams corpus is three or four. Larger groups might benefit more from the proposed approaches.

6.3 CHAPTER SUMMARY

In this chapter, I utilized an existing short-term asymmetric measure, introduced in (Danescu-Niculescu-Mizil et al., 2011), to estimate pairwise entrainment and proposed new multi-party measures that better incorporate group dynamics. As the first step toward development of multi-party local entrainment, in the first part of this chapter, I investigated the hypothesis that local entrainment might occur within a time lag in groups. To test this hypothesis, I proposed a temporal window-based approach. The results of the three prediction tasks (process conflict, favorable outcome, and fake vs real) do not support my hypothesis. There is no significant improvement as a result of increasing time-lag. There are some non-significant improvements but the results are not consistent across all these tasks.

In the second part, I proposed a graph-based vector representation for multi-party entrainment to better incorporate the dynamics of entrainment relations, estimated by a pairwise asymmetric measure. I proposed three approaches to learn the multi-party entrainment vector representations from entrainment graphs. I performed multiple experiments to find the best setting for each of these methods. I concluded that the graph embedding learned by the proposed weakly-supervised approach outperforms the baselines and the other proposed approaches. Beside performance, this approach has other advantages. First, a vector is potentially more informative than a single-valued entrainment measure since each dimension is defined to capture a distinctive property of the entrainment graph which encodes the strength and dynamics of entrainment behaviors in the group. Moreover, it is computationally less expensive than the best performing baseline.

7.0 INCORPORATING INDIVIDUAL AND TEAM LEVEL ATTRIBUTES: GENDER

Work on dyads and groups has found that gender is related to both the strength and utility of entrainment (Levitan and Hirschberg, 2011; Yu and Litman, 2019). In the Teams Corpus, minimum convergence was higher for teams with greater gender diversity than teams with less gender diversity (Yu and Litman, 2019). In the Columbia Games Corpus, acoustic-prosodic entrainment was strongest for mixed-gender pairs, followed by female pairs, followed by male pairs; however, even though less prevalent, entrainment was more important to the dialogue success of the male pairs (Levitan and Hirschberg, 2011). The effects of gender composition on team performance are mixed, but in general, diverse teams have worse processes and performance than homogeneous teams (Mannix and Neale, 2005; Joshi and Roh, 2009; van Knippenberg and Schippers, 2007). More important is when gender diversity reflects deeper diversity such as differences in attitudes and schemas (Harrison et al., 1998). In this chapter, building upon the existing work that found relation between gender and entrainment, I attempt to answer the question that how can we incorporate gender or gender-composition of teams to develop enhanced measures of multi-party entrainment? In this chapter, I only focus on gender and gender composition. But, any relevant individual or team-level factor, such as prior game experience, age, and education could be incorporated by the proposed approaches.

I propose two approaches. The first approach is an extension to the Hierarchical Alignment Model (HAM) (Doyle and Frank, 2016) by utilizing a crossed hierarchy design and grouping data by gender-composition variable beside marker or team variable. The second approach is an extension to the proposed entrainment vector representation learning model, discussed in section 6.2, by jointly learning to predict the gender or gender-composition.

7.1 GENDER-AWARE HIERARCHICAL ALIGNMENT MODEL (GHAM)

The HAM model (Doyle et al., 2016b) which is the strongest baseline in previous section is shown in Figure 7.1. Similar to the method (Danescu-Niculescu-Mizil et al., 2011) explained in section 6.1.1, the HAM model measures entrainment on a set of pairwise exchanges between the speakers and estimates the conditional likelihood of an event occurring (presence of a marker category) in the utterances. For more details, on how I build the pairwise utterance sets and how I calculate the likelihoods see section 6.1.1. The alignment in Ham model is defined in the log-odd space:

$$HAM(b,a) \approx logit^{-1}p(e_b|e_a) - logit^{-1}p(e_b|\neg e_a) \tag{7.1}$$

HAM assumes that word use in replies is shaped by whether the preceding message contained the marker of interest. The binomial probability μ is dependent on whether the preceding message did (μ_{align}) or did not (μ_{base}) contain the marker, and the inferred alignment value is the difference between these probabilities in log-odds space (η_{align}). Marker token counts $C^{align}_{m,a,b}$ and $C^{base}_{m,a,b}$ are draws from a binomial distribution. $N^{base}_{m,a,b}$ and $N^{base}_{m,a,b}$ are the total token counts. The remainder of the model is a hierarchy of normal distributions. There are three levels in the hierarchy: marker or lexical category level, conversational subgroup or team level, and conversational dyad level. All of the normal distributions have identical standard deviations $\sigma^2 = 0.25$. A $Uniform[-5,5]$ distribution gives a relatively uninformative prior for the baseline marker frequency. The normal distribution prior for the alignment hierarchy is centered at zero.

I define two gender-aware HAM models using two of the most obvious extensions to the original HAM architecture [1]. First, utilizing a crossed hierarchy design and a dyad-level gender-composition index that indicates whether the pair is same-gender or mixed-gender, I model pair-level entrainment η where $\eta_{m,c}$ represents the deviation from the average for each gender-composition and on each marker category. The graphical model is in Figure 7.1. C here is a binary variable that indicates if the pair is same-gender (male-male and female-

[1] I experimented with other variations of these architectures and choices that we could make about the hyper-priors or the gender-composition variable (binary, or different categorical versions). I did not get any improvement and I do not report them here.

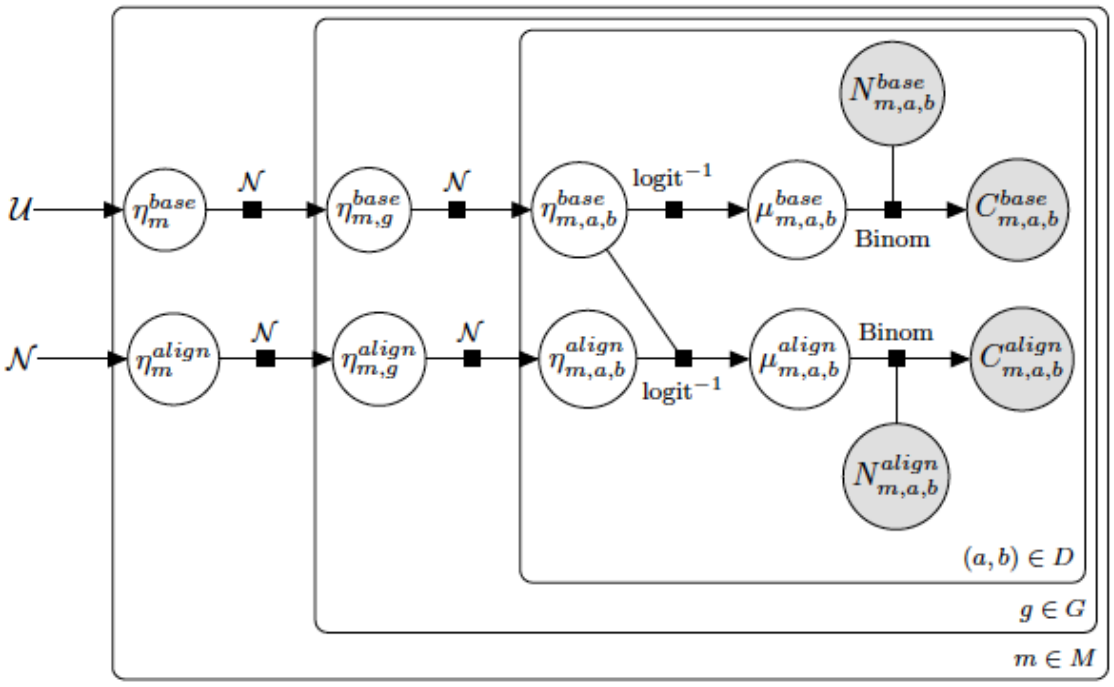
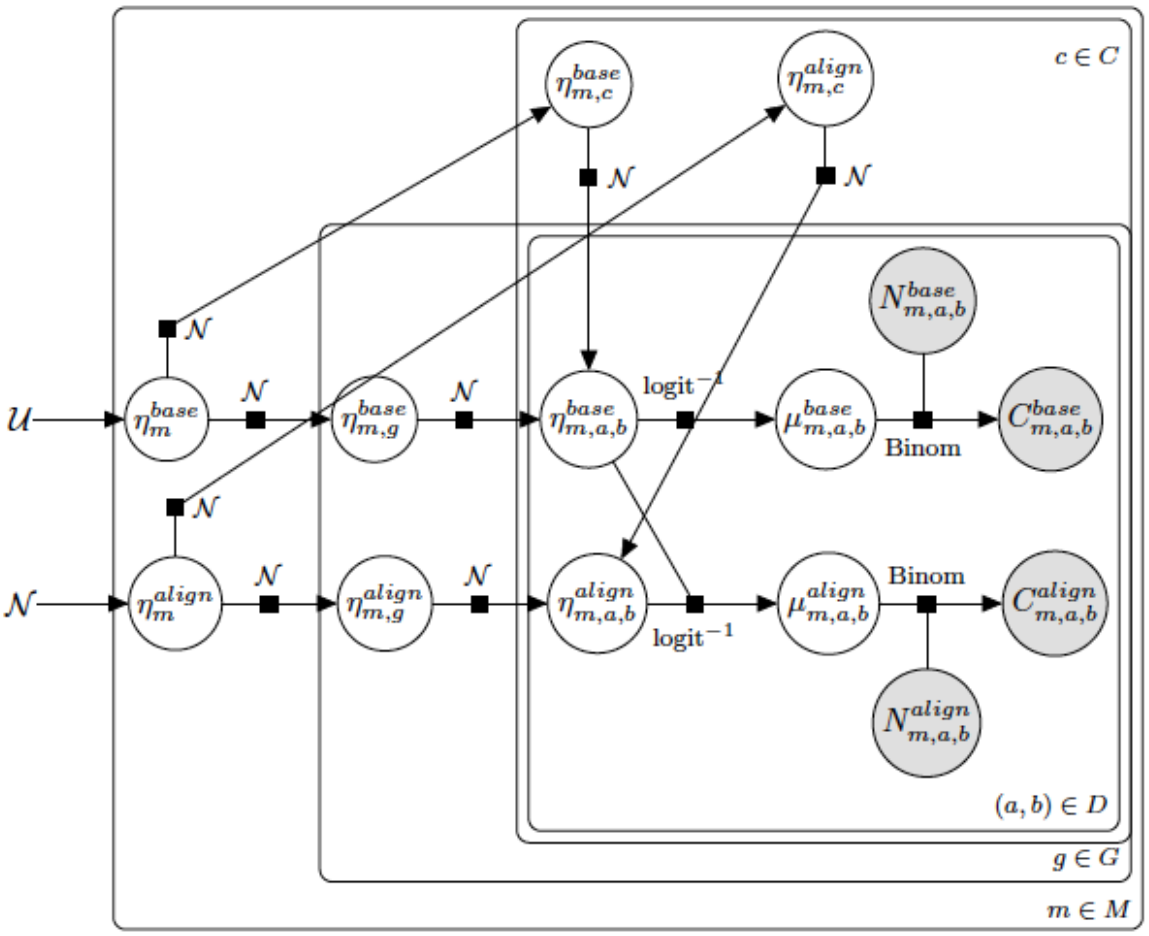
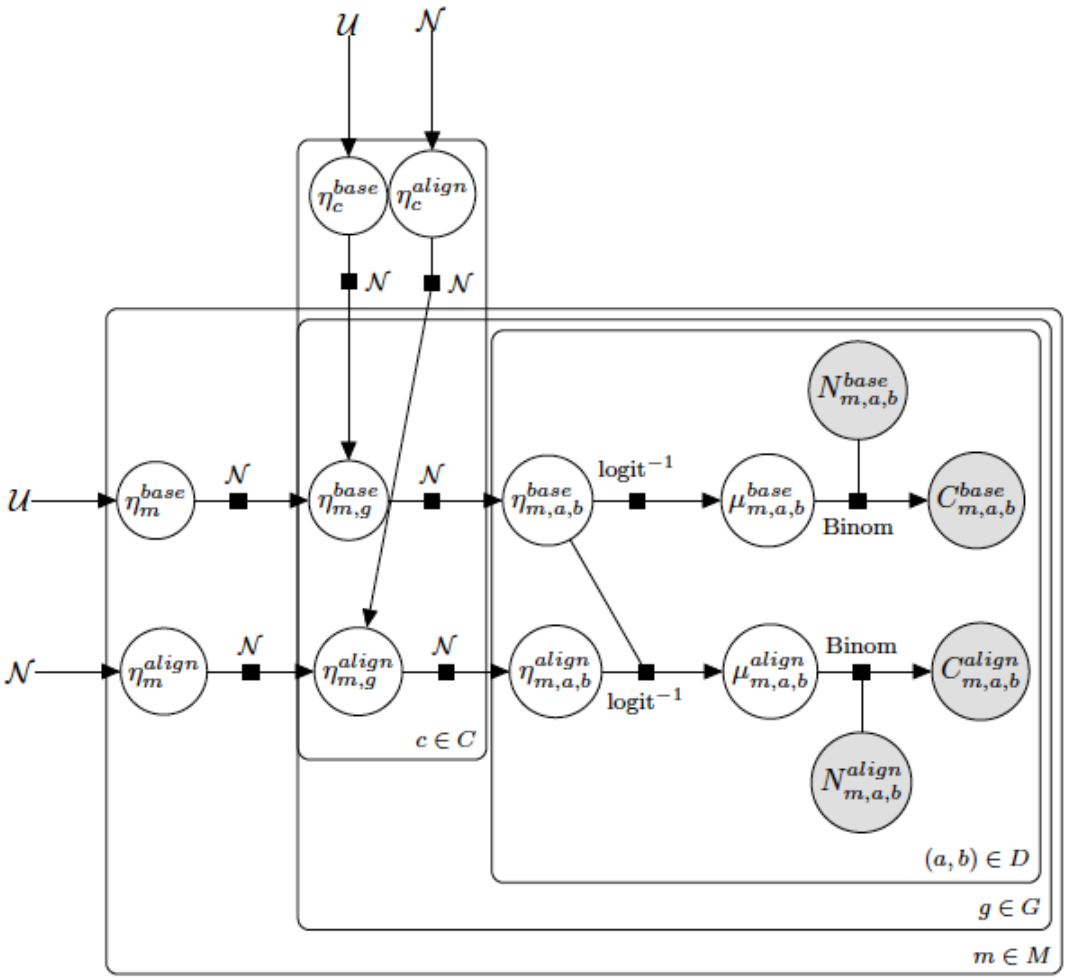


Figure 9: Graphical model of Hierarchical Alignment Model (HAM). G represent the teams, D is dyads, and M is lexical markers.


$c \in C$
$\eta^{base}_{m,c}$
$\eta^{align}_{m,c}$
$\mathcal{N}$
$\mathcal{N}$
$N^{base}_{m,a,b}$
$\mathcal{U}$
η^{base}_{m}
$\mathcal{N}$
$\eta^{base}_{m,g}$
$\mathcal{N}$
$\eta^{base}_{m,a,b}$
logit^{-1}
$\mu^{base}_{m,a,b}$
Binom
$C^{base}_{m,a,b}$
$\mathcal{N}$
$\mathcal{N}$
η^{align}_{m}
$\mathcal{N}$
$\eta^{align}_{m,g}$
$\mathcal{N}$
$\eta^{align}_{m,a,b}$
logit^{-1}
$\mu^{align}_{m,a,b}$
Binom
$C^{align}_{m,a,b}$
$N^{align}_{m,a,b}$
$(a,b) \in D$
$g \in G$
$m \in M$


$\mathcal{U}$
$\mathcal{N}$
η_c^{base}
η_c^{align}
$\mathcal{N}$
$\mathcal{N}$
$N_{m,a,b}^{base}$
$\mathcal{U}$
η_m^{base}
$\mathcal{N}$
$\eta_{m,g}^{base}$
$\mathcal{N}$
$\eta_{m,a,b}^{base}$
logit^{-1}
$\mu_{m,a,b}^{base}$
Binom
$C_{m,a,b}^{base}$
$\mathcal{N}$
η_m^{align}
$\mathcal{N}$
$\eta_{m,g}^{align}$
$\mathcal{N}$
$\eta_{m,a,b}^{align}$
logit^{-1}
$\mu_{m,a,b}^{align}$
Binom
$C_{m,a,b}^{align}$
$N_{m,a,b}^{align}$
$c \in C$
$(a,b) \in D$
$g \in G$
$m \in M$

female) or mixed-gender. Second, I utilize a crossed hierarchy design and a binary team-level gender-composition index. In this model, I define gender-composition as a grouping on the team-level entrainment variables for all the marker categories with hyper-priors the same as hyper-priors of the marker level [2]. The graphical model is in Figure 7.1. C here is a binary variable that indicates if the team has equal or more males, or it has more females.

I used PyStan to implement these models. The optimization algorithm is Hamiltonian Monte Carlo (HMC). The model is fit with 2000 iterations of the sampler (1000 as warm-up) and four chains. For the second model, I increased the values of the hyper-parameters (max_treedepth=15,adapt_delta=0.9) so the model would converge. I extracted entrainment estimates ($\eta_{m,g}$) from each of the final 1000 iterations of the model, and I report upper-bound, lower-bound, and mean of the 95% highest posterior density interval on the parameter values.

7.2 VECTOR REPRESENTATION LEARNING: MULTI-TASKING

Multi-Task Learning (MTL) has been used successfully in natural language processing (Collobert and Weston, 2008; Deng et al., 2013) and computer vision (Zhang et al., 2014). MTL comes in many forms and serves different goals. The goal of MTL in (Caruana, 1997) is defined as "improving generalization by leveraging the domain-specific information contained in the training signals of related tasks". Auxiliary tasks in MTL help to improve the main task by biasing the model to prefer representations that other tasks also prefer.

Given the prior studies that have found relations between gender and entrainment (Levitan and Hirschberg, 2011), I propose to extend the proposed entrainment vector representation learning models by multi-task learning. The main task is one of the weakly supervised or self-supervised approaches proposed in previous section. The auxiliary task is to predict the gender of speaker or the gender-composition of the team.

The Embedding layer is shared between the two tasks. The main task has a output layer which is a sigmoid of dot product of node vector and graph vector. The network structure

[2] I experimented with nested versions of gender-composition and marker but the model did not converge. So, I do not report them here.

of the main task is in Figure 8. The second task has two fully connected dense layers on top of the shared embedding layer. The first dense layer's input dimension is 10 (the input is a graph or node vector) with sigmoid activation function and has 20 units. The second dense layer or the output layer has a single unit and a sigmoid activation function. I apply dropouts on the embedding and the first dense layer with dropout rate of 20% for regularization.

The challenge in our problem is that the main task and the auxiliary task do not have the exact same inputs. We can only predict gender for the real nodes and graphs (not the fake ones). To solve this issue, I define a custom loss function for the auxiliary task. The custom loss function is a masked binary cross entropy, given that the output label is binary, which has the ability to mask unlabeled non-relevant permuted data. the total loss is the sum of the losses from the main task and the auxiliary task.

The input of the main task is pairs of node and graph indexes. I define two different auxiliary tasks. The first defined auxiliary task gets the node index as the input and predicts if the input node is for a male or a female speaker. The second auxiliary task gets the graph index as the input and predicts the gender-composition of the input graph. For simplicity, I define gender composition as a binary variable which indicates if the team has more female members or less or equal female members.

7.3 EXPERIMENTS AND RESULTS

I evaluate the quality of the proposed gender-aware approaches on the independent classification task of predicting social outcomes. I use the multi-party entrainment values as features to predict social outcomes. I use support vector machines with RBF kernel and perform leave-one-out cross validation. I have 119 instances. For more details see section 6.2.3.1. The results are in Table 16. The only gender-aware extension that improves the performance of it's non gender-aware approach is multi-tasking on the weakly-supervised approach where the auxiliary task is to predict the binary gender composition of the teams (graphs). This method could improve the performance of the predicting social outcome from 63.86 to 66.38. This is the best result that I have on this prediction task overall. The results of the same

Features	conflict	conflict+FS(Lasso)	favaroble	favaroble+FS(Lasso)
HAM	**68.90**	69.74	53.78	51.26
GHAM-dyad	66.38	63.02	51.26	51.26
GHAM-team	68.90	68.90	51.26	52.10
WeaklyS_KP_T	58.82	63.86	56.30	57.14
MTL_WeaklyS_KP_T(Node)	63.02	67.22	59.66 2	57.14
MTL_WeaklyS_KP_T(Graph)	59.66	64.70	62.18	60.50
WeaklyS_KP_NT	59.66	**74.79**	63.86	63.86
MTL_WeaklyS_KP_NT(Graph)	65.54	67.22	**66.38**	**66.38**
Self_K_P	63.02	73.95	57.14	58.82
MTL_Self_K_P(Graph)	57.14	60.50	56.30	60.50

Table 16: Accuracy of binary classification of process conflict and favorable outcome using proposed gender-aware models. (Node) or (Graph) indicates whether the auxiliary task was to predict gender using nodes, or gender-composition using graphs. The bold numbers are the best result in each column.

model on prediction of conflict is also improved but the original weakly supervised approach performs better when incorporating feature selection. None of the other extensions could improve the original methods.

The results of the multi-tasking on the weakly-supervised approach using team-level factor of gender-composition are promising. This is a first step toward the study of team-level factors when measuring multi-party entrainment. I chose a very simple network for the auxiliary task, did not perform parameter tuning, and did not experiment with other team-level factors, such as mean or standard deviation of age or education level.

7.4 CHAPTER SUMMARY

In this chapter, building upon the existing work that found relation between gender and entrainment, I attempted to answer the question that how can we incorporate gender or gender-composition of teams to develop enhanced measures of multi-party entrainment? The proposed model to extend the entrainment vector representation learning by multi-task learning shows some promising results and motivates further investigations in future.

8.0 CONCLUSION AND FUTURE WORK

8.1 SUMMARY OF CONTRIBUTIONS AND RESULTS

This research has five major contributions. First, I studied the relationships of entrainment at two linguistic levels, acoustic-prosodic and lexical, in multi-party dialogues. I found that first, entrainment occurs at both levels. Second, entrainment at these linguistic levels positively correlate (i.e., teams that entrain on one level are more likely to entrain on the other, and vice versa). Finally, to predict positive and negative team outcomes, a multimodal model with features from acoustic-prosodic and lexical levels outperforms a unimodal model.

Second, I introduced a new weighted convergence measure for multi-party entrainment at the global setting which utilizes group convergence dynamics to weight pair convergences. Experimental results show that the proposed weighted measure has higher validity than simply averaging convergence in that it is more predictive of fake versus real conversations. Also, the proposed weighted measure has higher utility at the downstream task of predicting the favorable team outcome. In general, a proper weighting of the dyad-level convergences performs better than non-weighted averaging in multiple tasks.

Third, at the short-term (local) setting, where similarity is measured between dyadic exchanges, I investigated the effect of considering time-lag between source and target utterances using a temporal window approach. The source utterance is preceding to the target and is a potential source of entrainment for it. The results of the evaluation experiments do not support the hypothesis that considering time-lag improves the multi-party local entrainment measure. There is no significant improvement as a result of increasing time-lag. There are some non-significant improvements, but the results are not consistent across all three evaluation tasks.

Fourth, I proposed a novel graph-based vector representation for multi-party local entrainment by incorporating strength and direction of pairwise entrainment relations. The proposed kernel approach and weakly-supervised representation learning method show promising results at the downstream task of predicting team outcomes. Also, the vector representation is potentially more informative than the single-valued baselines.

Finally, I proposed new methods to employ non-conversational factors such as gender-composition to entrainment measures. The proposed multi-task vector representation learning, where predicting gender-composition is an auxiliary task, shows promising improvement at predicting the favorable outcome. The results indicate this is a promising path that requires further investigation in future.

Table 17 summarizes the hypotheses I made in this thesis and if the evaluation results support these hypotheses.

8.2 TRADEOFFS

I developed several measures of multi-party entrainment in this work. As I discussed before, each speaker might show one or multiple forms of entrainment on one or multiple linguistic features. So, we need to estimate entrainment using different measures such as global convergence and local entrainment to fully understand the entrainment behavior of the speakers. Although, each of the introduced measures might be more proper for a certain situation.

Convergence is a within team measure. It is directly estimated using the given team's data. It does not include any learning process and does not require training data. So, it can be used both online and offline. Although, estimating lexical convergence on short dialogues or on small intervals have sparsity issue.

Proximity compares the target group with non-team members. Similar to convergence, proximity is directly estimated and does not involve any learning process and does not require any training data. But, it does require data from other teams on the same task, same topic of discussion, and similar audio recording situation to be compared to the target team as a sort of baseline. This measure can be used both online and offline if the baseline data is

Hypothesis	Supported?
H1: Entrainment occurs at multiple linguistic modalities during conversation.	Yes
H2: Teams that entrain on one modality are more likely to entrain on the other, and vice versa.	Yes
H3: To predict team outcomes, a multi-modal model with entrainment features from acoustic-prosodic and lexical levels outperforms a unimodal model.	Yes
H4: A properly weighted convergence measure has a higher validity than the simple-averaging convergence.	Yes
H5: A properly weighted convergence measure has a higher utility than the simple-averaging convergence at the downstream task of predicting team outcomes.	Yes
H6: Time-lagged multi-party entrainment estimated by the proposed window-based approach has a higher validity than the immediately preceding baseline.	No
H7: Time-lagged multi-party entrainment estimated by the proposed window-based approach has a higher utility than the immediately preceding baseline at the downstream task of predicting team outcomes.	No
H8: The proposed vector representation of multi-party entrainment outperforms the state of the art adaptive entrainment baselines at the downstream task of predicting team outcomes.	Partially Supported
H9: The proposed gender-aware extension of the entrainment measures outperforms the none-extended versions at the downstream task of predicting team outcomes.	Partially Supported

Table 17: The thesis hypotheses and summary of the final conclusions.

collected beforehand.

The proposed entrainment embedding utilizes the pairwise measure introduced by Danescu-Niculescu-Mizil et al. (2012). This pairwise measure is more suitable for lexical entrainment. Any other directional pairwise measure can be utilized to build the entrainment graphs. For example, for acoustic-prosodic features, we can utilize the PCA-based directional entrainment measure (Lee et al., 2011).

The entrainment embedding estimated directly using the kernel method does not involve any learning process and does not require any training data. It can be used both online and offline. The self-supervised and weakly-supervised approaches are more proper for offline situations since they involve training time to learn the embedding. Although the weakly-supervised approach is computationally more expensive than the kernel approach, it is superior to the kernel approach in terms of accuracy at predicting team outcomes. It should be emphasized that the training data, utilized to learn the embedding, does not require any manual labeling.

All these measures can be utilized for multiple applications. One example application is to understand team interactions. Researchers might introduce team interventions for less entraining teams to optimize the team processes. These measures can also be used to design dialogue systems with entrainment abilities for multi-party situations.

8.3 LIMITATIONS AND FUTURE WORK

This study utilizes **one single dataset**, the teams corpus, which is a **small** dataset collected in a **controlled laboratory experiment**. The teams are **small groups** of three or four people. Each participant was a member of only a single team. Also, the team outcomes are self-reported **perceived** variables. Here I describe several future directions that arise from this thesis and are yet to be investigated:

- **Real Data**: This work is done using the Teams Corpus which is collected in a laboratory setting. The laboratory setting enabled high-quality audio and video capture, while the experimental study allowed us to collect measures of team processes. Although this data

approximates task-oriented multi-party spoken dialogues, future work should experiment with real world situations to both validate our results and provide solutions for new challenges like noisy input or multiple topics of discussion.

- **Utility of the Proposed Measures**: In this thesis, utility of the proposed measures is verified at the downstream task of predicting perceived team outcomes. Using other datasets, the utility of the proposed entrainment measures should also be verified at predicting real, not-perceived outcomes such as task success [1].
- **Varying Team Size**: This study deals with groups of three or four people. Future work should experiment with groups with varying team-size to both validate our results on large groups and provide solutions for any potential challenges that arise when the teams are big. For example, although I did not find any evidence that considering time-lag improves the local entrainment measure, this conclusion might not hold for larger groups.
- **Multi-party Dialogue Systems**: An interesting end application of this research is entrainment-enabled multi-party dialogue systems. A few studies have implemented such systems for two-party dialogues (Lubold et al., 2015; Lopes et al., 2015; Hu et al., 2016; Mizukami et al., 2016). Future work should attempt to implement entrainment-enabled dialogue systems for multi-party enviroments, such as computer-assisted collaborative learning, or home-bots with ability to detect multiple users.
- **Relation of Different Entrainment Measures**: In this work, I examined multiple entrainment measures. Each of these measures focuses on a different aspect of entrainment. But, I did not investigate the relation between these measures. Future work should investigate these relations. For example, what is the relation between convergence and proximity, or global and local measures in multi-party groups? Do teams with high proximity also have high convergence? Is there any relation between global convergence and local entrainment?
- **An Exhaustive Multi-Modal Analysis**: In this work, I investigated the relation between entrainment at acoustic-prosodic and lexical levels in multi-party spoken dialogues.

[1] In the Teams corpus, almost all of the teams successfully finish the task. So, we could not employ a task success measure in this thesis using the Teams corpus.

Future work should extend this study by including more linguistic levels such as syntax or semantics, and other modalities such as visual (e.g., gesture and facial expressions). Following this work, Levitan et al. (2018) investigated the relation of acoustic-prosodic and lexical entrainment on two corpora of dyadic dialogues and they did not find any correlations. Future work should investigate these relations on other multi-party corpora.

- **Incorporating Individual and Team-level Non-Conversational Characteristics**: In this thesis, I attempted to enhance multi-party entrainment measures by incorporating gender and teams' gender-composition. This work takes the first step and shows this is a promising path to follow. Other non-conversational characteristics, such as age, education, and familiarity with the task should be investigated in future work.
- **Optimizing the Hyper-Parameters and Finding the Best Network Architectures**: The goal of this research is not to find the best prediction models and the best possible result for the proposed models. I introduce several novel ideas and research paths for multi-party entrainment. The promising results using most obvious network structures and intuitively chosen parameters indicates the potential of these novel ideas. Future work should investigate finding the best results by optimizing the hyper-parameters and by investigating other network structures.

APPENDIX A

TRANSCRIPTION OF A DIALOGUE OF A 3-PERSON TEAM PLAYING THE GAME IN THEIR FIRST SESSION FROM THE TEAMS CORPUS

Engineer : Ok I'm going to

Engineer : shore up these two.

Pilot : Good move.

Engineer : Then we got one and then I guess I can also

Engineer : Can I use my powers twice in one play

Messenger : Mm

Pilot : yes

Engineer : OK well I guess yeah cause we (–)

Pilot : Well the Pilot's limited to once per turn.

Messenger : yeah

Engineer : Ok and then I have (–) two treasure cards.

Messenger : two treasure cards

Pilot : Two Treasure cards.

Messenger : Mmhmm

Engineer : yeooh

Engineer : Ok so, let me see sorry

Pilot : uh oh

Engineer : so i move this up one tick

Pilot : and then these are gonna get shuffled

Engineer : Mmhmm

Pilot : If we had any sandbags we'd wanna use them while these were getting shuffled.

Engineer : (–) shuffle and then discard this into the treasure

Engineer : Um and I guess I have to, I'm still picking two cards I guess then.

Pilot : yeah

Messenger : Mmhmm

Pilot : That's the worst part. It starts out easier but then it starts sinking faster and faster.

Engineer : So Iron Gate and Phanthom Rock.

Messenger : Iron Gate

Messenger : umm

Pilot : That's the Iron Gate

Messenger : Ok

Pilot : Phathom Rock

Messenger : Phathon Rock, oh um

Engineer : Aw, oh my gosh.

Pilot : Oh sinking already

Engineer : Ok

Messenger : I guess the Phathom Rock

Pilot : Alright. The Phathom Rock Card gets removed

Engineer : Oh yeah. That sucks.

Pilot : Foruntately it's not critical. No Treasures on it and it's not a path to anywhere.

Messenger : Mmhmm

Pilot : Ok that was your turn. So

Engineer : Ok

Pilot : I can take up to three actions.

Pilot : (–)

Pilot : I'm gonna move one.

Pilot : Opps that not sunken, why am I turning it.

Pilot : I move two ,

Pilot : give me two and two that doesn't help alright.

Pilot : I'm just gonna move two and sit there and draw

Pilot : one

Pilot : Draw two.

Engineer : Hmm.

Pilot : Draw Flood Cards equal to the water level.

Messenger : Lost.

Pilot : Lost Lagoon

Engineer : Oh.

Pilot : And Whispering Gardens.

Pilot : Alright that's it for my turn.

Messenger : um hmm.

Messenger : Ok so.

Messenger : I think I'm going to...

Messenger : use my power to fly here and save this card.

Pilot : Ok.

Pilot : One action left.

Messenger : Um.

Messenger : then.

Messenger : Move here.

Messenger : And then

Pilot : Wait isn't using your power an action?

Messenger : move

Messenger : Yeah one action.

Pilot : So you one action to move, one to shore up, and then one move is what I'm asking.

Messenger : Um, (sighs) there's not much I can do, ok

Pilot : third action

Pilot : Not at this point. We need to get matching sets and get treasures before the island sinks.

Messenger : (it's) another Helicopter.

Messenger : that and then two Flood Cards.

Pilot : alright

Pilot : alright

Messenger : Um Cliffs of Abandon and Breaker's Bridge.

Pilot : Breaker's Bridge.

Messenger : and Cliffs of Abandon

Engineer : Ok first i'm going to

Engineer : use (–)

Engineer : huh (–)

Engineer : Um I'm gonna (–) Cliffs of Abandon

Engineer : and

Engineer : and then I'll take two Treasure cards ok.

Messenger : Two cards

Engineer : Thank you.

Engineer : OK

Pilot : (you got it)

Engineer : Wait op, six, I gotta discard one.

Pilot : Nope.

Pilot : You've got five.

Engineer : Oh wait (–) sorry (–).

Messenger : Oh isn't it five?

Pilot : You got five and you got three fire each. You need to get to

Pilot : Cave of Shadows

Engineer : Yeah ok um so Temple of the Moon and Golden Gate.

Messenger : Umm.

Pilot : Gold Gate is removed. That's not good.

Engineer : Mm, that's oh.

Messenger : Temple of the Moon.

Pilot : Temple of the Moon. Ah.

Messenger : Oh and the Golden Gate card.

Engineer : Oh yeah I'm sorry.

Pilot : Take that SanFrancisco

Pilot : Alright it's your turn, so it's my turn.

Pilot : Um, I'm gonna shore up the Temple of the Moon.

Pilot : I got two of these. I kinda wanna stay there right now. Nothin' else I can shore up so I think I'll only take one action.

Pilot : I'm gonna go to six cards though. so I'm gonna go ahead and use Sandbags so

Pilot : shore that up

Pilot : Oo Water's Rise that's not good.

Pilot : tick's up one.

Messenger : Oh did you only take one Treasure card?

Pilot : M cause I had to play the Water's Rise immediately but

Pilot : Also got a Lion.

Pilot : Shuffle shuffle shuffle shuffle shuffle shuffle

Pilot : And I'm going three

Pilot : Breaker's Bridge. Good thing I shored Temple of the Moon up.

Messenger : Breaker's Bridge

Pilot : Temples of the Moon and Whispering Gardens.

Messenger : Whispering Garden's keeps getting turned over.

Engineer : Yeah

Pilot : Good news is they keep on floating right next to the Engineer

Pilot : Ok that's it for my turn.

Messenger : Ok.

Messenger : Um so.

Messenger : Ah yeah five cards.

Engineer : I can discard one if

Messenger : Ok.

Pilot : Yeah well you gotta, you gotta be adjacent to her now so you'd have to fly there and then hand her

Messenger : Mmk. I'm going to fly

Messenger : here.

Pilot : or is it adjacent or is it on the same

Messenger : is it on the same?

Pilot : same island yeah you'd have to fly to meet her.

Messenger : Same island, ok here.

Messenger : I'm gonna give you this.

Engineer : ok.

Pilot : you gotta discard a treasure cards.

Engineer : ok um

Engineer : thanks.

Messenger : Ok so that's two turns so far.

Pilot : Yup, you got one left.

Messenger : Um.

Pilot : You could either move or shore up.

Messenger : I'm going to...

Messenger : You can shore up adjacent right?

Pilot : Yes.

Engineer : mmhmm

Messenger : So I mine as well just shore this one up.

Pilot : Yeah it's got the Treasure on it though.

Messenger : Ok and then two Treasure cards.

Messenger : Sandbag this one.

Messenger : And then two Flood cards.

Pilot : Three Flood cards.

Messenger : Three Flood cards.

Messenger : Cliffs of Abandon, Lost Lagoon, Iron Gate.

Engineer : Mmmm.

Messenger : Mmm Iron-

Engineer : Um soo

Engineer : Alright I'm going to move two places wait.

Engineer : Maybe that was waste for me (–)

Pilot : You could move two and collect a treasure.

Engineer : Yeah. I just (–) I'm gonna save something.

Engineer : Wait is collect treasure a move?

Pilot : Yes.

Engineer : Oh ok.

Engineer : One Two. And then I will- I guess I put these in the treasure room

Pilot : Yup and you collect the rasberry jello.

Messenger : And that was three turns?

Pilot : Yup.

Engineer : Yeah that's it. Thank you.

Messenger : Mmhmm.

Engineer : um yups got these guys and I guess three Flood cards unfortunately.

Engineer : What's it gonna be. Copper Gate, Crimson Forest and Howling garden.

Messenger : Howling Garden gone.

Pilot : There's copper gate. oop

Messenger : And what's the other one?

Engineer : Ah Crimson Forest.

Messenger : Oh.

Messenger : Mkay

Pilot : Alright, what am I doing here. So I have two of these. I'm gonna shore this up for one. And

Pilot : and then I'm gonna use my special power for one action messsenger this to you for one action messenger this for you. each got two. flip over two.

Messenger : Thank you

Messenger : Oh I have to discard. One, two, three.

Pilot : Well you could immediately use either Sandbags or Helicopter Lift.

Messenger : Oh um.

Pilot : Oh yeah I need definitely want to use it cause I just drew a Water's Rise.

Messenger : Use this to save that?

Engineer : Hhhmm.

Pilot : (sighs)

Messenger : Throw that out and now I have five (–).

Pilot : (–) up at three (–) shuffling.

Messenger : Could I also use this one?

Pilot : Yes you could. Just use it any time.

Messenger : Umm.

Messenger : I'm going to

Messenger : move myself away, just in case those fall away.

Pilot : If it falls away you'll still be able to swim but

Messenger : to here

Pilot : Alright.

Pilot : (fool)

Pilot : flip Breaker's Bridge, Howling Gardens and Temple of the Moon.

Messenger : Oh thank God we saved that one.

Engineer : Nope this is gone. and temple of the moon is sunk oh.

Messenger : Breaker's Bridge is Gone.

Messenger : Oh the Breaker's Bridge card.

Pilot : Oh yeah.

Messenger : ok um.

Pilot : Mm

Messenger : I'm going to give that to you for one action.

Pilot : You need to be on the same as me.

Messenger : Keep forgetting.

Messenger : Yo're up there.I don't think I could. I could..

Pilot : I'm up there. you could- you could use your Pilot ability to go there.

Messenger : Oh that's one.

Messenger : two.

Pilot : two.

Pilot : You got one left. You could shore it up.

Messenger : Umm.

Messenger : Shore that up.

Pilot : And that'll be three.

Pilot : Don't draw another Water's Rise.

Messenger : (–)

Messenger : Helicopter Lift and this.

Pilot : Three Lions.

Engineer : I'm so far away from everything.

Pilot : Alright then three Flood Cards.

Messenger : Oh.

Pilot : yup

Messenger : OK Crimson Forest, um Iron Gate.

Pilot : Op, both those are getting removed.

Messenger : And Cliffs of Abandon. uh

Engineer : Hmm.

Messenger : ok

Engineer : um oh this is a little tricky. I have so many cards to give and no one to give them to because I'm a million miles away.

Messenger : Oh um I can use Helicopter Lift and move you to our side.

Engineer : That would just be gorgeous so I can- ok and then so I guess that counts as one move for me.

Pilot : Mm. Helicopter Lift does not that's not an action so you still have all three. That's one, that's two.

Engineer : Oh ok. ok, so one and two. Um.

Messenger : Awesome.

Engineer : And then I am going to move back over here in hopes that- well I guess it would't matter but we could attempt at some point to save this ok. Um could I please have two?

Engineer : oh look at that ok.

Messenger : Awesome.

Engineer : Um now it's time for a bad part.

Pilot : now we just need to getcha there.

Engineer : Yeah.

Engineer : Ok um Whispering Garden, Lost Lagoon and Copper Gate. so Whispering Garden

Pilot : oop

Messenger : Ah (–) Gate is gone.

Pilot : That's a problem. That's a real problem.

Messenger : Oh what's the other one?

Engineer : Um Lost Lagoon and oop

Pilot : we now don't have a path back to landing.

Messenger : oh and I only have one Helicopter card.

Engineer : Ooo we're fun.

Messenger : And I can move myself.

Pilot : You can fly to anyone, but we're gonna need to be helicopter lifted off that side of the island. that's the only way we are gonna get there and I'm gonna need another one.

Engineer : yeah

Pilot : to get off Fool's Landing.

Pilot : Ok so that was your turn.

Engineer : Yes.

Pilot : I gotta rock.

Messenger : Oh we're actually, I think we're good. Cause we only need one more.

Pilot : That's one. well we also need to get back to Fool's Landing and escape.

Messenger : Oh I can send you that, well hmm, we need one more Helicopter Lift.

Engineer : Hmm.

Engineer : Yeah. Oh we'll figure it out. cool enough though but I don't know

Pilot : One.

Pilot : Ah. And I can move but moving-

Pilot : If we're stay in the same place, one helicopter Lift can get us there but we'll still need one more and you also need to get treasure.

Engineer : Well.

Engineer : yeah I mean if you use your two moves like to go here and he- wait ah sorry.

Pilot : No I want to stay with you and use one Helicopter Lift cause

Engineer : Well I'm saying she does like two moves to get here and she can flip this over and then she would only need one more.

Pilot : One to (–) ok

Engineer : Oh We that would be on like the next turn. Just to ensure it doesn't sink that one.

Pilot : Well

Pilot : that's

Pilot : one

Pilot : two.

Pilot : And flipping over three flood card

Engineer : I don't know.

Pilot : Cave of Embers

Pilot : Silver Gate

Messenger : Silver Gate

Pilot : Dunes of Deception.

Engineer : (–) we don't care about them

Pilot : Alright. I'm done.

Messenger : Mmkay um.

Messenger : Ah (–) this.

Messenger : I'm going to use my power to move, no this is me.

Pilot : Oh wait

Messenger : Wait yeah.

Messenger : Oh.

Pilot : you were blue?

Messenger : That sucks.

Messenger : I'm gonna use my powers.

Pilot : right Wow, that'll teach me to refer to the colors.

Messenger : to go

Pilot : Go to one of the Lion spaces.

Messenger : eehh.

Messenger : Umm.

Messenger : I think I want to go over here.

Messenger : So that's one.

Pilot : One (–)

Messenger : Use another one to flip it over.

Messenger : oh Wait can I still take it even thought it's flipped over?

Pilot : Mm.

Pilot : Yes.

Messenger : yeah so never mind I'm not gonna flip it over.

Engineer : Mmm.

Messenger : Um.

Pilot : One.

Messenger : Ss, two to get the lion.

Engineer : ok

Pilot : two to get the lion.

Messenger : And then

Messenger : One more. Um

Messenger : Hmm.

Pilot : Either shore up or move yourself one closer to the Fool's Landing.

Messenger : Um.

Messenger : I'll just shore it up.

Pilot : Alright draw two Treasure cards hope the waters don't rise.

Messenger : Nope. Waters went rising.

Messenger : And then three floods

Pilot : three flood card.

Messenger : ok.

Messenger : Bronze Gate.

Messenger : Ah Coral Palace and Tidal Palace.

Engineer : Ok, umWould it even matter (where) we move?

Pilot : we wanna be on the same place so we can be Helicopter lifted off of here.

Engineer : Ok um

Pilot : And you could shore up two before you move.

Engineer : yeah I mean

Pilot : Oh.

Engineer : I think only this one is adjacent to me so.

Pilot : Right cause you can't diagonal.

Messenger : Mmhmm.

Engineer : so I'll shore up this one.

Engineer : Then I'll come over here, um.

Engineer : I don't, I won't use that (–) yeah

Messenger : Just give the-

Messenger : yup, water rises.

Engineer : Ok water rises, but also helicopter lift with water rises so I have to put this up one tick, shuffle those and put them back in the flood thing

Engineer : and then you have to immediately discard those. thinkin'

Engineer : Um, yeah I can

Pilot : (–)

Engineer : I can do, I can use a Helicopter Lift and then you can use yours to get there finally so

Pilot : Right.

Pilot : Right get to the two of us to Fool's Landing with the helicopter lift.

Engineer : yeah.

Engineer : (–) treasure cards

Pilot : And then I think we've got it.

Pilot : Play it out.

Pilot : Bronze Gate.

Messenger : Bronze Gate

Pilot : Opp I'm flipping all three over.

Messenger : Oop, that one's out. Bronze Gate's out.

Messenger : Oh.

Pilot : Dunes of Deception is out.

Engineer : um

Messenger : And Howling Garden.

Engineer : howlin-

Pilot : Howling Garden.

Pilot : Ok at this point I'm waitin' for the Helicopter Lift out of here and we've got three.

Pilot : I can shore this up as an action.

Pilot : and I can sit here and twidle my thumbs as an action.

Pilot : Flip over two.

Pilot : Two.

Pilot : and flip over

Pilot : well we've lost the Lost Lagaoon, imagine that.

Messenger : Lost Lagoon.

Pilot : Did we fi- uh was it already removed?

Engineer : oh I think we did

Pilot : And (–)

Engineer : (–)

Pilot : Silver Gate.

Pilot : Whispering Garden.

Engineer : Whispering Garden's gone

Pilot : Temple of the Moon.

Engineer : We were really lucky.

Engineer : Pilot

Messenger : I seriously thought it was this one.

Messenger : Ok um my turn?

Pilot : Yup. Need your pilot ability to get there.

Messenger : Use my power to go over there.

Messenger : And use helicopter lift to get us all out.

Pilot : (–) high fives all around.

Engineer : This is good. Yay!

Engineer : (laughs)

Messenger : (laughs)

APPENDIX B

THE SUBSET OF THE QUESTIONNAIRES RELATED TO THE FAVORABLE OUTCOME AND CONFLICT

Q36 Thinking of your team, please choose the letter A through F that best matches for each item.

- A: There is not a friendly atmosphere among people. (1)
- B (2)
- C (3)
- D (4)
- E (5)
- F: There is a friendly atmosphere among people. (6)

Q27 Thinking of your team, please choose the letter A through F that best matches for each item.

- A: People in my group do not trust each other. (1)
- B (2)
- C (3)
- D (4)
- E (5)
- F: People in my group trust each other (6)

Q22 Please use the following scale to rate your agreement on each item.

	1: Highly Inaccurate	2	3	4	5: Highly Accurate
Working together energizes and uplifts members of our team.	❍	❍	❍	❍	❍
There is a lot of unpleasantness among members of this team.	❍	❍	❍	❍	❍
The longer we work together as a team, the less we do.	❍	❍	❍	❍	❍
Every time sometime attempts to correct a team member whose behavior is not acceptable, things seem to get worse rather than better.	❍	❍	❍	❍	❍
My relations with other team members are strained.	❍	❍	❍	❍	❍
I very much enjoy talking and working with my teammates.	❍	❍	❍	❍	❍
The chance to get to know my teammates is one of the best parts of working on this team.	❍	❍	❍	❍	❍

Q23 Please use the following scale to rate your agreement on each item.

	1: Strongly disagree	2	3	4	5: Strongly agree
I enjoy the kind of work we do in this team.	❍	❍	❍	❍	❍
Working on this team is an exercise in frustration.	❍	❍	❍	❍	❍
Generally speaking, I am very satisfied with this team.	❍	❍	❍	❍	❍

Q28 Thinking of your team, please choose the letter A through F that best matches for each item.

- A: People are not warm and friendly. (1)
- B (2)

- C (3)
- D (4)
- E (5)
- F: People are warm and friendly. (6)

Q30 Thinking of your team, please choose the letter A through F that best matches for each item.

- A: People do not treat each other with respect. (1)
- B (2)
- C (3)
- D (4)
- E (5)
- F: People treat each other with respect. (6)

Q31 Thinking of your team, please choose the letter A through F that best matches for each item.

- A: People do not work well together as a team. (1)
- B (2)
- C (3)
- D (4)
- E (5)
- F: People work well together as a team. (6)

Q32 Thinking of your team, please choose the letter A through F that best matches for each item.

- A: People do not cooperate with each other. (1)
- B (2)
- C (3)
- D (4)
- E (5)
- F: People cooperate with each other. (6)

Q33 Thinking of your team, please choose the letter A through F that best matches for each item.

- A: People are not willing to share resources. (1)
- B (2)
- C (3)
- D (4)
- E (5)
- F: People are willing to share resources. (6)

Q34 Thinking of your team, please choose the letter A through F that best matches for each item.

- A: People almost never speak well of the group. (1)
- B (2)
- C (3)
- D (4)
- E (5)
- F: People almost always speak well of the group. (6)

Q35 Thinking of your team, please choose the letter A through F that best matches for each item.

- A: The people are not proud to belong to the group. (1)
- B (2)
- C (3)
- D (4)
- E (5)
- F: The people are proud to belong to the group. (6)

Q37 Please choose the number from 1 to 5 that fits best, from 1 (to no extent) to 5 (to a great extent).

	1: To no extent (1)	2: To a limited extent (2)	3: To some extent (3)	4: To a considerable extent (4)	5: To a great extent (5)
This team has confidence in itself.	❍	❍	❍	❍	❍
This team believes it can become unusually good at its tasks.	❍	❍	❍	❍	❍
This team expects to be a high-performing team.	❍	❍	❍	❍	❍
This team feels it can solve any problem it encounters.	❍	❍	❍	❍	❍
This team believes it can be very productive.	❍	❍	❍	❍	❍
This team can get a lot done when it works hard.	❍	❍	❍	❍	❍
No task is too tough for this team.	❍	❍	❍	❍	❍

Q38 Please use the following scale to rate your agreement on each item.

	Strongly Disagree (1)	Disagree (2)	Neither Agree nor Disagree (3)	Agree (4)	Strongly Agree (5)
In my group, we have similar thoughts about the best way to proceed.	❍	❍	❍	❍	❍
In my group, we eventually agree on what to do.	❍	❍	❍	❍	❍
In my group, we have similar ideas about how to go about winning the game.	❍	❍	❍	❍	❍

Q38 Please use the 1 to 5 scale to answer the questions below. Please choose the number that fits best.

	1: None (1)	2 (2)	3 (3)	4 (4)	5: A lot (5)
How frequently did you have disagreements within your group about the task you were working on?	❍	❍	❍	❍	❍
How often were there disagreements about who should do what in your group?	❍	❍	❍	❍	❍
How much relationship tension was there in your group?	❍	❍	❍	❍	❍
How much conflict of ideas was there in your group?	❍	❍	❍	❍	❍
How much conflict was there in your group about task responsibilities?	❍	❍	❍	❍	❍
How often did people get angry while working in your group?	❍	❍	❍	❍	❍
How often did people in your group have conflicting opinions about the task you were working on?	❍	❍	❍	❍	❍
How often did you disagree about resource allocation in your group?	❍	❍	❍	❍	❍
How much emotional conflict was there in your group?	❍	❍	❍	❍	❍

BIBLIOGRAPHY

Alderisio, F., Fiore, G., Salesse, R. N., Bardy, B. G., and Di Bernardo, M. (2017). Interaction patterns and individual dynamics shape the way we move in synchrony. *Scientific reports*, 7(1):6846.

Beal, D. J., Cohen, R. R., Burke, M. J., and McLendon, C. L. (2003). Cohesion and performance in groups: A meta-analytic clarification of construct relations. *Journal of Applied Psychology*, 88:989–1004.

Beňuš, Š. (2014). Conversational entrainment in the use of discourse markers. In *Recent Advances of Neural Network Models and Applications*, pages 345–352. Springer.

Beňuš, Š., Gravano, A., Levitan, R., Levitan, S. I., Willson, L., and Hirschberg, J. (2014). Entrainment, dominance and alliance in supreme court hearings. *Knowledge-Based Systems*, 71:3–14.

Beňuš, Š., Levitan, R., and Hirschberg, J. (2012). Entrainment in spontaneous speech: the case of filled pauses in supreme court hearings. In *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*, pages 793–797. IEEE.

Beňuš, Š., Trnka, M., Kuric, E., Marták, L., Gravano, A., Hirschberg, J., and Levitan, R. (2018). Prosodic entrainment and trust in human-computer interaction. pages 220–224.

Boersma, P. and Heuven, V. v. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

Borrie, S. A., Lubold, N., and Pon-Barry, H. (2015). Disordered speech disrupts conversational entrainment: a study of acoustic-prosodic entrainment and communicative success in populations with communication challenges. *Frontiers in psychology*, 6.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177.

Branigan, H. P., Pickering, M. J., and Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13 – B25.

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, 96:41–44.

Brennan, S. E. and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Chartrand, T. L. and Bargh, J. A. (1999). The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Coulston, R., Oviatt, S., and Darves, C. (2002). Amplitude convergence in children's conversational speech with animated personas. In *Seventh International Conference on Spoken Language Processing*.

Danescu-Niculescu-Mizil, C., Gamon, M., and Dumais, S. (2011). Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., and Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, pages 699–708.

de Kok, I., Heylen, D., and Morency, L.-P. (2013). Speaker-adaptive multimodal prediction model for listener responses. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 51–58. ACM.

De Looze, C., Scherer, S., Vaughan, B., and Cambpell, N. (2014). Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*, 58:11–34.

Deng, L., Hinton, G., and Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8599–8603. IEEE.

Doyle, G. and Frank, M. C. (2016). Investigating the sources of linguistic alignment in conversation. In *ACL (1)*.

Doyle, G., Goldberg, A., Srivastava, S. B., Frank, M. C., et al. (2016a). Alignment at work: Accommodation and enculturation in corporate communication. Technical report.

Doyle, G., Yurovsky, D., and Frank, M. C. (2016b). A robust framework for estimating linguistic alignment in twitter conversations. In *Proceedings of the 25th international conference on world wide web*, pages 637–648. International World Wide Web Conferences Steering Committee.

Driskell, J. E., Salas, E., and Hughes, S. (2010). Collective orientation and team performance: Development of an individual differences measure. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52:316–328.

Finkelstein, S., Scherer, S., Ogan, A., Morency, L.-P., and Cassell, J. (2012). Investigating the influence of virtual peers as dialect models on students' prosodic inventory. In *Third Workshop on Child, Computer and Interaction*.

Ford, L. R. and Fulkerson, D. R. (2009). Maximal flow through a network. In *Classic papers in combinatorics*, pages 243–248. Springer.

Friedberg, H., Litman, D., and Paletz, S. B. F. (2012). Lexical entrainment and success in student engineering groups. In *Proceedings Fourth IEEE Workshop on Spoken Language Technology (SLT)*, Miami, Florida.

Friedkin, N. E. and Johnsen, E. C. (2011). *Social influence network theory: A sociological examination of small group dynamics*, volume 33. Cambridge University Press.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., and Tylén, K. (2012). Coming to terms: quantifying the benefits of linguistic coordination. *Psychological science*, 23(8):931–939.

Gevers, J. M., Rutte, C. G., and Van Eerde, W. (2006). Meeting deadlines in work groups: Implicit and explicit mechanisms. *Applied psychology*, 55(1):52–72.

Giles, H., Coupland, N., and Coupland, I. (1991). 1. accommodation theory: Communication, context, and. *Contexts of accommodation: Developments in applied sociolinguistics*, 1.

Gonzales, A. L., Hancock, J. T., and Pennebaker, J. W. (2010). Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37:3–19.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM.

Guzzo, R. A., Yost, P. R., Campbell, R. J., and Shea, G. P. (1993). Potency in groups: Articulating a construct. *British Journal of Social Psychology*, 32:87–106.

Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*.

Harrison, D. A., Price, K. H., and Bell, M. P. (1998). Beyond relational demography: Time and the effects of surface- and deep-level diversity on work group cohesion. *Academy of Management Journal*, 41:95–107.

Heath, J. S. (2017). *Causes and Consequences of Convergence*. PhD thesis, University of California at Berkeley.

Hu, Z., Halberg, G., Jimenez, C. R., and Walker, M. A. (2016). Entrainment in pedestrian direction giving: How many kinds of entrainment? In *Situated Dialog in Speech-Based Human-Computer Interaction*, pages 151–164. Springer.

Hung, H., Huang, Y., Friedland, G., and Gatica-Perez, D. (2011). Estimating dominance in multi-party meetings using speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):847–860.

Iwata, T. and Watanabe, S. (2013). Influence relation estimation based on lexical entrainment in conversation. *Speech Communication*, 55(2):329–339.

Jain, M., McDonough, J., Gweon, G., Raj, B., and Rosé, C. P. (2012a). An unsupervised dynamic bayesian network approach to measuring speech style accommodation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 787–797. Association for Computational Linguistics.

Jain, M., McDonough, J. W., Gweon, G., Raj, B., and Ros, C. P. (2012b). An unsupervised dynamic bayesian network approach to measuring speech style accommodation. In *EACL*, pages 787–797.

Jehn, K. A. and Mannix, E. A. (2001). The dynamic nature of conflict: A longitudinal study of intragroup conflict and group performance. *Academy of management journal*, 44(2):238–251.

John, O. P., Donahue, E. M., and Kentle, R. L. (1991). The big five inventory-versions 4a and 54. University of California, Berkeley, Institute of Personality and Social Research. http://www.ocf.berkeley.edu/ johnlab/bfi.htm.

Joshi, A. and Roh, H. (2009). The role of context in work team diversity research: A meta-analytic review. *Academy of Management Journal*, 52:599–627.

Jovanovic, N., op den Akker, R., and Nijholt, A. (2006). Addressee identification in face-to-face meetings. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Lee, C.-C., Katsamanis, A., Black, M. P., Baucom, B. R., Georgiou, P. G., and Narayanan, S. (2011). An analysis of pca-based vocal entrainment measures in married couples' affective spoken interactions. In *INTERSPEECH*, pages 3101–3104.

Levitan, R., Gravano, A., and Hirschberg, J. (2011). Entrainment in speech preceding backchannels. In *Proceedings of ACL/HLT*.

Levitan, R., Gravano, A., Willson, L., Benus, S., Hirschberg, J., and Nenkova, A. (2012). Acoustic-prosodic entrainment and social behavior. In *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–19.

Levitan, R. and Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Interspeech*.

Levitan, S. I., Xiang, J., and Hirschberg, J. (2018). Acoustic-prosodic and lexical entrainment in deceptive dialogue. In *Proc. 9th International Conference on Speech Prosody 2018*, pages 532–536.

Lin, C.-Y. and Hovy, E. (2000). The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 495–501. Association for Computational Linguistics.

Litman, D., Paletz, S., Rahimi, Z., Allegretti, S., and Rice, C. (2016). The teams corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431.

Lopes, J., Eskenazi, M., and Trancoso, I. (2015). From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech & Language*, 31(1):87–112.

Lubold, N. and Pon-Barry, H. (2014). Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 5–12. ACM.

Lubold, N., Pon-Barry, H., and Walker, E. (2015). Naturalness and rapport in a pitch adaptive learning companion. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 103–110. IEEE.

Luo, Y., Wang, Q., Wang, B., and Guo, L. (2015). Context-dependent knowledge graph embedding. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1656–1661.

Mannix, E. and Neale, M. A. (2005). What differences make a difference? the promise and reality of diverse teams in organizations. *Psychological Science in the Public Interest*, 6:31–55.

Matarazzo, J. D. and Wiens, A. N. (1967). Interviewer influence on durations of interviewee silence. *Journal of Experimental Research in Personality*.

Matsuyama, Y. and Kobayashi, T. (2015). Towards a computational model of small group facilitation. In *2015 AAAI spring symposium series*.

Metzing, C. and Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49(2):201–213.

Michalsky, J., Schoormann, H., and Niebuhr, O. (2018). Conversational quality is affected by and reflected in prosodic entrainment. In *Proceeding the 9th International Conference on Speech Prosody 2018*, pages 389–392.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Miron, E., Erez, M., and Naveh, E. (2004). Do personal characteristics and cultural values that promote innovation, quality, and efficiency compete or complement each other? *Journal of Organizational Behavior*, 25:175–199.

Mitchell, C. M., Boyer, K. E., and Lester, J. C. (2012). From strangers to partners: Examining convergence within a longitudinal study of task-oriented dialogue. In *SIGDIAL Conference*, pages 94–98.

Mizukami, M., Yoshino, K., Neubig, G., Traum, D., and Nakamura, S. (2016). Analyzing the effect of entrainment on dialogue acts. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 310–318.

Mukherjee, A. and Liu, B. (2012). Analysis of linguistic style accommodation in online debates. In *Proceedings of COLING 2012*, pages 1831–1846.

Mullen, B. and Copper, C. (1994). The relation between group cohesiveness and performance: An integration. *Psychological Bulletin*, 115(2):210–227.

Munson, S. A., Kervin, K., and Robert Jr, L. P. (2014). Monitoring email to indicate project team performance and mutual attraction. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 542–549. ACM.

Nadeau, C. and Bengio, Y. (2000). Inference for the generalization error. In *Advances in neural information processing systems*, pages 307–313.

Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32(5):790.

Nenkova, A., Gravano, A., and Hirschberg, J. (2008). 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*, HLT-Short '08, pages 169–172.

Niederhoffer, K. G. and Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Oertel, C. and Salvi, G. (2013). A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 99–106. ACM.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Porzel, R., Scheffler, A., and Malaka, R. (2006). How entrainment increases dialogical effectiveness. In *Proceedings of the IUI'06 Workshop on Effective Multimodal Dialogue Interaction*, pages 35–42.

Rahimi, Z., Kumar, A., Litman, D., Paletz, S., and Yu, M. (2017). Entrainment in multi-party spoken dialogues at multiple linguistic levels. *Proc. Interspeech 2017*, pages 1696–1700.

Rahimi, Z. and Litman, D. (2018). Weighting model based on group dynamics to measure convergence in multi-party dialogue. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 385–390.

Rahimi, Z., Litman, D., and Paletz, S. (2019). Acoustic-prosodic entrainment in multi-party spoken dialogues: Does simple averaging extend existing pair measures properly? In *Advanced Social Interaction with Agents*, pages 169–177. Springer.

Reichel, U. D., Beňuš, Š., and Mády, K. (2018). Entrainment profiles: Comparison by gender, role, and feature set. *Speech Communication*, 100:46–57.

Reitter, D. and Moore, J. D. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28.

Reitter, D. and Moore, J. D. (2007). Predicting success in dialogue. In *Proceedings of the 45th Meeting of the Association of Computational Linguistics*, pages 808–815.

Richardson, D. C., Dale, R., and Tomlinson, J. M. (2009). Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science*, 33(8):1468–1482.

Roche, J. M., Dale, R., and Caucci, G. M. (2012). Doubling up on double meanings: Pragmatic alignment. *Language and Cognitive Processes*, 27(1):1–24.

Romero, D. M., Galuba, W., Asur, S., and Huberman, B. A. (2011). Influence and passivity in social media. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 18–33. Springer.

Shin, H. and Doyle, G. (2018). Alignment, acceptance, and rejection of group identities in online political discourse. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 1–8.

Street Jr, R. L. (1984). Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research*, 11(2):139–169.

Sun, C. and Morency, L.-P. (2011). Towards speaker adaptation for dialogue act recognition. *Ron Artstein, Mark Core, David DeVault, Kallirroi Georgila, Elsi Kaiser, and Amanda Stent*, page 188.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077. International World Wide Web Conferences Steering Committee.

Tang, J., Sun, J., Wang, C., and Yang, Z. (2009). Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816. ACM.

Thomason, J., Nguyen, H. V., and Litman, D. (2013). Prosodic entrainment and tutoring dialogue success. In *16th International Conference on Artificial Intelligence in Education*, pages 750–753, Memphis, TN.

Truong, K. P. and Heylen, D. (2012). Measuring prosodic alignment in cooperative task-based conversations. In *Thirteenth Annual Conference of the International Speech Communication Association*.

van Knippenberg, D. and Schippers, M. C. (2007). Work group diversity. *Annual Review of Psychology*, 58:515–541.

Wageman, R., Hackman, J. R., and Lehman, E. (2005). Team diagnostic survey: Development of an instrument. *The Journal of Applied Behavioral Science*, 41(4):373–398.

Wang, Y., Reitter, D., and Yen, J. (2014). Linguistic adaptation in conversation threads: Analyzing alignment in online health communities. In *Proceedings of the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics*, pages 55–62.

Wang, Y., Reitter, D., and Yen, J. (2017). How emotional support and informational support relate to linguistic alignment. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 25–34. Springer.

Wang, Y., Yen, J., and Reitter, D. (2015). Pragmatic alignment on social support type in health forum conversations. *Proc. Cognitive Modeling and Computational Linguistics (CMCL). Association for Computational Linguistics, Denver, CO*, pages 9–18.

Ward, A. and Litman, D. (2007a). Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In *SLaTE Workshop on Speech and Language Technology in Education (ISCA Tutorial and Research Workshop)*, Farmington, PA.

Ward, A. and Litman, D. (2007b). Measuring convergence and priming in tutorial dialog. *University of Pittsburgh*.

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

Wendt, H., Euwema, M. C., and van Emmerik, I. H. (2009). Leadership and team cohesiveness across cultures. *The Leadership Quarterly*, 20(3):358–370.

Xing, W. and Ghorbani, A. (2004). Weighted pagerank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pages 305–314. IEEE.

Xu, Y., Cole, J., and Reitter, D. (2018). Not that much power: Linguistic alignment is influenced more by low-level linguistic features rather than social power. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 601–610.

Yu, M. and Litman, D. (2019). Investigating the relationship between multi-party linguistic entrainment, team characteristics and perception of team social outcomes. In *32nd International FLAIRS Conference*.

Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2014). Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer.

Zhou, C., Liu, Y., Liu, X., Liu, Z., and Gao, J. (2017). Scalable graph embedding for asymmetric proximity. In *Thirty-First AAAI Conference on Artificial Intelligence*.