The Teams Corpus and Entrainment in Multi-Party Spoken Dialogues

Diane Litman
University of Pittsburgh
Pittsburgh, PA

Susannah Paletz
University of Maryland
College Park, MD

Zahra Rahimi and **Stefani Allegretti** and **Caitlin Rice**
University of Pittsburgh
Pittsburgh, PA

Abstract

When interacting individuals *entrain*, they begin to speak more like each other. To support research on entrainment in cooperative multi-party dialogues, we have created a corpus where teams of three or four speakers play two rounds of a cooperative board game. We describe the experimental design and technical infrastructure used to collect our corpus, which consists of audio, video, transcriptions, and questionnaire data for 63 teams (47 hours of audio). We illustrate the use of our corpus as a novel resource for studying team entrainment by 1) developing and evaluating team-level acoustic-prosodic entrainment measures that extend existing dyad measures, and 2) investigating relationships between team entrainment and participation dominance.

1 Introduction

Linguistic entrainment[1] refers to the convergence of (para)linguistic features across speakers during conversation (Brennan and Clark, 1996; Porzel et al., 2006). Research has found that speakers entrain to both human and computer conversational partners, with the amount of entrainment often positively related to conversational and task success. However, most prior work has focused on the study of entrainment during two-party dialogues, rather than during the multi-party conversations typical of teams.

To support the study of entrainment during multi-party cooperative dialogue, we have created a large-scale corpus (over 47 hours of recordings) of teams of three or four speakers playing a cooperative board game requiring conversation. The corpus consists of audio, video, transcriptions, and questionnaire data for 63 teams. The goal of the corpus is to provide a freely-available data resource for the development and evaluation of multi-party entrainment measures that can be 1) computed using language technologies, 2) motivated and validated by the literature on teams, and 3) associated with measures of task and dialogue success.

In this paper, we first describe the experimental design and technical infrastructure used to create our corpus. We then present two case studies illustrating the use of our corpus as a novel resource for studying team entrainment: quantifying acoustic-prosodic entrainment at the team-level rather than the dyad-level, and incorporating a construct from the teamwork literature into the study of entrainment.

2 Background and Related Work

The development of methods for automatically quantifying entrainment in text and speech data is an active research area, as entrainment has been shown to correlate with success measures or with social variables for a variety of phenomena, e.g., acoustic-prosodic, lexical, and syntactic (Nenkova et al., 2008; Reitter and Moore, 2007; Mitchell et al., 2012; Levitan et al., 2012; Lee et al., 2011; Stoyanchev and Stent, 2009; Lopes et al., 2013; Lubold and Pon-Barry, 2014; Moon et al., 2014; Sinha and Cassell, 2015). Such research, in turn, requires corpora with certain properties. A high-quality spoken language corpus for studying entrain-

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

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1421–1431,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

ment would include transcriptions suitable for natural language processing, audio recordings suitable for signal processing, and meta-data such as task success or speaker demographics.

While most research has focused on quantifying the amount of entrainment between pairs of speakers, recent work has 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). To date, however, 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, computer-mediated discussions, lexical analysis of transcriptions). In this paper we both investigate group acoustic-prosodic entrainment and examine relationships between group entrainment and a factor from the teamwork literature called participation equality / dominance (Paletz and Schunn, 2011).

Also, while freely available speech corpora have supported the study of entrainment in two-party dialogues (e.g., Switchboard, Maptask, the Columbia Games Corpus, Let's Go), few community resources exist for the study of multi-party entrainment. Some multi-party resources are only text-based (e.g., the online Slashdot forum (Allen et al., 2014), chat dialogues (Afantenos et al., 2015)). Those speech resources that do exist are often less than ideal as they were created for other purposes (e.g., Supreme Court arguments (Beňuš et al., 2014; Danescu-Niculescu-Mizil et al., 2012), the AMI meeting corpus (Carletta et al., 2006)). Although not created to study entrainment, the KTH-Idiap Group-Interviewing corpus (Oertel et al., 2014) is perhaps most relevant as it was explicitly designed to support research on group dynamics. However, the corpus contains only 5 hours of speech, and participants were PhD students so did not differ on variables such as age and social status.

The Teams corpus presented and used in this paper was designed to add several notable extensions to existing multi-party spoken dialogue resources. In particular, the Teams corpus was experimentally collected to constrain the team processes, tasks, and outcomes in ways that facilitate an investigation of team entrainment. First, the corpus consists of over 45 hours of cooperative task-oriented dialogues between three or four speakers, where audio and video files were collected and transcribed using best practices for computational processing. Second, the corpus was collected using an experimental manipulation informed by the organizational and social psychological literature on team processes in order to create high versus low-entrainment conditions. Third, since the social psychological literature suggests that team dynamics are more complex than an average of dyadic interactions, validated questionnaires were used to collect relevant variables of interest to researchers on teams, and individual participants were recruited so that teams would exhibit diversity with respect to these variables.

3 Experimental Study

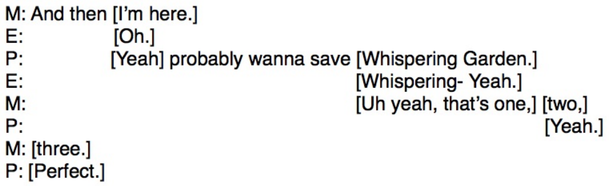
The Teams corpus was collected in a laboratory experiment. The laboratory setting enabled high-quality audio and video capture, while the experimental study allowed manipulations to vary entrainment and to collect measures of team processes.[2]

3.1 Design

Our data collection was via an experiment with a 2 by 2 within-and-between subjects design. Teams of 3-4 participants spent 2-3 hours in our lab taking self-report questionnaires and being audio and video-taped playing a cooperative board game. Two manipulations were designed to increase the likelihood of task success and entrainment [3]. For the first manipulation, half the teams were given a *teamwork* training intervention in which participants were given specific advice based on a needs analysis of the team skills important to the game (Gregory et al., 2013). Such mixed teamwork/taskwork training has been shown to improve team process outcomes (Salas et al., 2008). The other half only had

[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; Levitan et al., 2011). Our corpus is approximately 5 times larger, includes speech from teams rather than from dyads, and relatedly includes new types of team-related meta-data. Our corpus also contains both video and audio as our dialogues were face-to-face rather than restricted to voice.

[3] As discussed in Section 2, prior research has often found positive relationships between success and entrainment.



M: And then [I'm here.]
E: [Oh.]
P: [Yeah] probably wanna save [Whispering Garden.]
E: [Whispering- Yeah.]
M: [Uh yeah, that's one,] [two,]
P: [Yeah.]
M: [three.]
P: [Perfect.]

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.

training on the rules of the game, which all teams received.

For the second manipulation, each team played two isomorphic versions of the game. The game was originally designed to be played multiple times, with each session unique depending on the random placement of specific board tiles and the order of deck cards. To maintain experimental control, two specific deck card orders and board tile patterns that had the same underlying opportunities and obstacles were created. 33 teams played one game first, and 30 teams played the other game first. In either case, by the second time, the team should have a better grasp of the game and appropriate strategies.

3.2 Task

For the team task, we chose the cooperative board game Forbidden Island^{TM}, where players take on the roles of adventurers seeking treasures on an island before it is flooded. We chose this game because it both demands collaboration and is logistically feasible for our experiment. 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), lending itself directly to eliciting entrainment. Further, the game gives each player a different role to achieve the team goals, as well as game-specific terminology, generalizing to real-world situations with teamwork (e.g. aviation, health care). Logistically, Forbidden Island^{TM} can be played equally well with three or four players. This feature allowed us to schedule teams of four participants, but still play the game even if only three showed up. A typical game is also short enough to be played twice within an experimental session. Game rules were adapted to ensure the game difficulty was suitable for novice players (e.g., requiring three rather than four treasures be found before completing the game). As noted in Section 3.1, 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.

3.3 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), we advertised in non-student locations in predominantly ethnic minority neighborhoods.

3.4 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. Teams in the intervention condition (the between-subjects manipulation of our experimental design) were given an extra 10 minutes before the first game to receive training about teamwork strategies such as team roles, communication needs, and how to coordinate their actions (Gregory et al., 2013), as well as additional information adapted for the Forbidden Island^{TM} task itself. 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). The intervention condition teams were also given an additional 5 minutes before the second game to discuss what went well and poorly with their team processes. Finally, both between and after the two games, all participants filled out questionnaires regarding their team processes.

3.5 Data Capture

Game participants were located around a round table 48 inches in diameter in our 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 multichannel 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 videos provide information about the games that are not always obvious from the audio[4], as well as non-verbal data for future analysis (e.g., of gesture or posture).

4 The Teams Corpus

Our experiment ran from February through August 2015, yielding over 47 hours of recordings from 63 teams[5] (216 individuals).

4.1 Descriptive Statistics

	Control (n=31)		Intervention (n=32)	
	3-per.	4-per.	3-per.	4-per.
# of teams	20	11	16	16
avg g1 time	26.6	28.0	26.4	27.3
avg g2 time	18.0	17.7	18.2	19.7

Table 1: Team descriptives ($n = 63$).

The 216 participants in our 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). Thus, our recruitment yielded demographically diverse participants in ways that are useful for team research.

Table 1 shows the distribution of the teams in our corpus by experimental condition (control versus intervention) and 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 3-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 experimental condition ($p > .7$) or by team size ($p > .3$), and there were also no interaction effects.

Our team-level data provides preliminary evidence for the success of one of our experimental manipulations, as second games were significantly shorter than first games. [6]

4.2 Audio Segmentation and Transcription

After the experiment was completed, our multiple audio track speech was manually segmented and transcribed using the Higgins Annotation Tool[7].

[4] We are currently using the videos to annotate game-specific measures of task success.

[5] A power analysis for our experiment yielded a minimum target sample size of 52 teams.

[6] The time to complete a game is an easy to compute but a shallow (inverse) success measure. We are currently annotating our data for game-specific and dialogue-based success measures, and will also examine success in terms of team process measures computable from the questionnaires (Section 4.3).

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

Each audio track, which corresponds to each individual player, appears on a separate line in Higgins. A time stamp line applies to all of the (synchronized) audio tracks. 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 our 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. Within each transcription, text segments may also be defined and assigned values. We are using segments to annotate non-lexical aspects such as laughs.

4.3 Questionnaire Data

The pre-game questionnaire was used to collect individual demographic information such as discussed in Section 4.1, 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 was collected as 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).

4.4 Public Release

The Teams corpus will be freely available for research purposes[8], with the first release coordinated with the publication of this paper. The team level contents of the first release will consist of 63 game 1 and 62 [9] game 2 WAV files. The individual level contents of this release will consist of the demographic responses for the 216 participants in XLSX format. Later corpus releases will include associated audio segmentations and transcriptions in XML format, game-level video files, and personality and team process measures.

5 Case Studies Using the Teams Corpus

This section presents results from two case studies illustrating the use of the Teams corpus for novel research in multi-party dialogue entrainment. The first study proposes new team level measures that build on existing dyad-level measures of proximity and convergence, then uses these team measures to investigate whether prior dyad-level acoustic-prosodic entrainment findings generalize to teams. The second study investigates relationships between team convergence and participation equality / dominance.

5.1 Acoustic-Prosodic Team Entrainment

Speakers do not entrain on all linguistic features of conversations, and when they do entrain, they may entrain in different ways on different features. In this section we examine whether teams entrain on different acoustic-prosodic features during each of their two game conversations. Our current approach to measuring team-level entrainment is based on averaging dyad-level measures. We build on two dyad measures, namely, proximity and convergence (Levitan and Hirschberg, 2011). In a conversation, proximity measures feature similarity over the entire conversation, while convergence measures an increase in feature proximity over time.

5.1.1 Feature Extraction from Speaker Audio

We focus on the acoustic-prosodic dimensions of pitch, intensity, and voice quality, following previous work on dyad entrainment (Levitan and Hirschberg, 2011; Lubold and Pon-Barry, 2014; Borrie et al., 2015). Pitch is related to the frequency of the sound wave. Intensity describes the rate of energy flow. Jitter and shimmer are measures of variations of frequency and energy, respectively, which are descriptive of voice quality. We use the Praat software (Boersma and Heuven, 2002) to extract the following 9 acoustic-prosodic features: minimum (min), maximum (max), mean and standard deviation (SD) of pitch; min, max, mean of intensity;

[8] https://sites.google.com/site/teamentrainmentstudy/corpus

[9] One audio file was not properly saved during the experiment. The corresponding single-channel audio extracted from the game's video will be provided instead.

local jitter[10]; and local shimmer[11]. Features are extracted separately for each speaker and for each game. Before feature extraction, each game-level audio file for each speaker is pre-processed to remove silences (using a threshold of 1 second).

5.1.2 Measuring Team Proximity

Proximity quantifies the similarity of a feature value between conversational partners over their entire conversation. Intuitively, if a team has entrained on a feature in terms of proximity during a particular game, speakers within the same team should be more similar (or equivalently, less different) to each other than to all the other speakers in the corpus who are not on their team and are playing the same game (i.e., game 1 or game 2). For each game we computed a team-level partner difference ($TDiff_p$) and a team-level other difference ($TDiff_o$). In Section 5.1.4 we report paired t-test analyses to infer entrainment within a game when $TDiff_p$ is significantly smaller than $TDiff_o$.

The partner difference for a speaker in a dyad (Levitan and Hirschberg, 2011) is the absolute difference between the feature value for a speaker and her partner. For each team, we averaged these absolute values for all members of the team:

$$TDiff_p = \frac{\sum_{\forall i \neq j \in team}(|speaker_i - speaker_j|)}{|team| * (|team| - 1)} \quad (1)$$

The other difference for a speaker in a dyad (Levitan and Hirschberg, 2011) is the mean of the absolute differences between the speaker's value for a feature and the values of each of the speakers in the corpus (for the same game number) with whom the speaker was not partnered (set X in Formula 2). For each team, we averaged these means for all the members of the team:

$$TDiff_o = \frac{\sum_{\forall i \in team}(\frac{\sum_j |speaker_i - X_j|}{|X|})}{|team|} \quad (2)$$

For proximity, all of the feature values were normalized within a game based on gender[12] using z-scores ($z = \frac{v_{ij} - \mu_j}{\sigma_j}$; v_{ij} = value of speaker i in game j where $j \in \{1, 2\}$, μ_j = gender mean in game j, and σ_j = gender standard deviation in game j.)

Feature	Game1	Game2
Pitch-min	0.844	0.193
Pitch-max	-1.092	0.022
Pitch-mean	-1.297	-1.294
Pitch-sd	-0.407	-1.652
Intensity-mean	-4.469*	-4.911*
Intensity-min	-2.653*	-2.069*
Intensity-max	-3.625*	-2.853*
Shimmer-local	-2.390*	-2.782*
Jitter-local	-1.242	-2.702*

Table 2: Proximity t-values of a paired t-test comparing team-level partner ($TDiff_p$) vs. other ($TDiff_o$). Negative t-values indicate that partner differences are smaller than other differences. * $p < .05$. $n = 62$.

5.1.3 Measuring Team Convergence

Intuitively, there is evidence of convergence when speakers within a conversation become more similar to each other later in the conversation. While feature value differences are compared across teams to infer proximity entrainment, feature value differences within a single team are compared across time for convergence entrainment. Since differing time intervals have been examined in the dyad literature, we compared features extracted from the first versus last three, five, and seven minutes of each game, as well as from the two game halves.[13] Convergence was inferred via paired t-tests when the partner differences (Equation 1) in the second time interval were significantly smaller than in the earlier time interval (e.g., the $TDiff_p$ in the last 3 minutes of game 1 is smaller than $TDiff_p$ in the first 3 minutes of game 1). To break the games into different time intervals for feature extraction, we used the raw audio files to extract the breaking points of the conversation and then mapped these points to each of the processed audio files where silence was removed.

5.1.4 Team-Level Entrainment Results

The proximity results are shown in Table 2. Negative t-values indicate that differences between speak-

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

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

[12] Normalization is done only for proximity, since comparisons for convergence are within (rather than between) teams.

[13] (Levitan and Hirschberg, 2011) also looked for convergence between the two halves of the first game in their corpus.

	First vs. last 3 minutes		First vs. last 5 minutes		First vs. last 7 minutes		First vs. second half	
Feature	Game1	Game2	Game1	Game2	Game1	Game2	Game1	Game2
Pitch-min	**2.474***	−0.709	1.487	−1.299	1.359	−1.622	0.329	−0.884
Pitch-max	**4.947***	1.260	1.892	−0.468	1.348	−0.424	0.457	0.627
Pitch-mean	−2.687*	0.109	−2.900*	0.417	−2.965*	−0.361	−1.905	−0.266
Pitch-sd	1.364	0.409	1.919	0.591	1.807	0.576	1.271	0.089
Intensity-mean	−0.275	−2.946*	−0.454	−2.245*	−0.229	−1.825	−0.360	−1.540
Intensity-min	0.595	−3.188*	−0.136	−4.335*	0.009	−3.317*	−0.972	−3.324*
Intensity-max	0.328	0.327	−0.731	1.081	−0.140	0.511	−0.222	0.469
Shimmer-local	**2.896***	−0.476	**3.396***	−1.941	**3.006***	−1.704	**2.794***	−0.914
Jitter-local	**3.205***	0.725	**2.796***	0.242	**2.867***	0.469	**2.973***	0.260

Table 3: Convergence t-values of paired t-tests comparing team-level partner differences ($TDiff_p$) of first 3, 5, 7 minutes vs. last 3, 5, 7 minutes, respectively, and of first vs. second game half, for each game. Positive t-values indicate convergence (i.e., that partner differences in the second interval are smaller than in the first). Negative t-values indicate divergence. Significant convergence results are in bold. * $p < .05$. $n = 62$.

ers who are all within the same team are smaller than differences between team members and other speakers in the corpus. Thus, negative values are indicative of team entrainment. The results show that the team members were significantly more similar to each other than to other speakers on intensity mean, min, and max and on shimmer for both games. Team-level entrainment on jitter was significant for only the second game.

The convergence results are shown in Table 3 for four different temporal comparison intervals. Comparison of the significant game 1 results shows that teams entrained on pitch min, pitch max, shimmer, and jitter in at least one of the intervals. Both shimmer and jitter converged for all choices of temporal units. For pitch, convergence was instead only seen using the first and last 3 minutes, which are the intervals farthest in the game from each other. The only feature that diverged during game 1 is pitch-mean. The rest of the features did not show significant team-level partner differences during game 1 for any temporal interval and thus exhibited maintenance, meaning that the team members neither converged nor diverged. During game 2, we observed maintenance for all features except for intensity-mean and intensity-min, which diverged. Together our results suggest that when teams in our corpus converged on a feature, they did so earlier in the experiment (namely, just during the first game, and sometimes just in the earliest part of the first game).

As a divergent validity check for convergence, for each of the 62 teams, we constructed artificial versions of the real conversations between team members: For each member of the team, we randomly permuted the silence and speech intervals extracted by Praat. Ideally, we should not see evidence of convergence within these constructed conversations. Our results confirm that there was no significant entrainment on either of the two constructed games, for all temporal comparison intervals and all features.

In summary, team acoustic-prosodic entrainment did not occur for all features. For the features that did show entrainment, results varied depending on whether proximity or convergence was examined, and by the time intervals compared. With respect to type of entrainment, when looking at the entire game 1, there was significant evidence of entrainment (proximity) on mean, min, max intensity, and shimmer. Although there was no significant proximity for min, max pitch and jitter, they did become more similar (converged) over time. With respect to time, team convergence was found for shimmer and jitter independently of temporal interval examined, but for pitch only when comparing the most distant temporal intervals in game 1.

5.2 Participation Equality / Dominance

Within psychology, equality of participation has been associated with successful team performance and decision-making (e.g., (Mesmer-Magnus and DeChurch, 2009; Stasser and Titus, 1987)). Within computational linguistics, balance of participation with respect to proposal of ideas was associated with more productive small group (online) conversations (Niculae and Danescu-Niculescu-Mizil, 2016).

Extending this literature, we perform a novel in-

	Model 1			Model 2		
	B	SEB	β	B	SEB	β
Session Length	0.187	0.067	0.328*	0.197	0.064	0.344*
Team Size	108.706	47.398	0.269*	69.721	47.980	0.173
Participation Dominance				−1077.747	429.130	−0.299*
Model R^2	0.186			0.266		
Model F	6.761*			7.015*		

Table 4: Summary of hierarchical regression analysis for variables predicting entrainment on pitch-max. * $p < .05$. $n = 62$.

vestigation of the association between participation equality/dominance and team entrainment, focusing on the time interval showing the most significant convergence results in Section 5.1.4 (entrainment on pitch-max, pitch-min, jitter, and shimmer from the first to last 3 minutes of game 1).

Equation 3 defines the participation of player i in a team, where $speech\_length_i$ is the sum of the lengths of the speech intervals of player i:

$$participation_i = \frac{speech\_length_i}{\sum_{m \in team} speech\_length_m} \quad (3)$$

Participation dominance in turn is the standard deviation of the participation for all team members:

$$\begin{aligned} Dominance &= \sigma(Participation), \\ Participation &= \{participation_i | i \in team\} \quad (4) \end{aligned}$$

Higher standard deviations indicate a greater range of participation from team members, and lower standard deviations indicate more participation equality.

We performed a hierarchical regression analysis for each of the four acoustic-prosodic features noted above as the target entrainment variable. As in the convergence section, we measured entrainment as the average differences ($TDiff_p$) of the team in the first interval minus the second interval. Larger positive numbers are indicative of more entrainment. The independent variables we included in our analysis are: team size, session length, average age of the team members, percentage of the female players in each team, and participation dominance. The first four are covariates that have been found to be or are likely related to team communication and/or dynamics. We hypothesized that participation dominance would be related to entrainment above and beyond these other potential variables.

Table 4 presents the results with pitch-max for entrainment. (The other 3 entrainment variables did not show significant relationships with participation.) The standardized βs indicate the effect size and direction of the individual variables on pitch-max, whereas the R^2 indicates the effect size of the model of all the variables together. Average age and percent female were not significantly related to entrainment on pitch-max, so were excluded from the final analyses.

First, both team size and session length were entered as potential independent variables into the regression analysis with pitch-max as the dependent variable. This model (Model 1) was significant. Specifically, team size and session length were both significantly positively associated with entrainment on pitch-max. That is, as team size or session length increased, entrainment also increased.

Participation dominance was then entered to create Model 2, which included team size, session length, and participation dominance. The amount of variance explained for participation dominance was significant above and beyond the variables entered in Model 1, $\Delta R^2 = 0.08$, $\Delta F(1, 58) = 6.307$, $p = 0.015$. Specifically, there was a significant negative association between participation dominance and entrainment on pitch-max, such that greater participation equality was related to greater entrainment. This suggests that the more each team member is given a chance to equally contribute, the more likely they are to entrain on their maximum pitch.

6 Summary and Broader Implications

The long-term goal of our research is to use speech and language processing, informed by the teamwork literature, to develop computational measures of conversational team entrainment that will be useful for predicting team success. We first described the design and contents of the Teams corpus, which is being made freely available for research purposes. Experimental manipulations, high-quality audio and video with time-aligned transcriptions, and self-reported team process data make the corpus a unique resource for studying multi-party dialogue entrainment. We provided two examples illustrating the use of the Teams corpus to facilitate new directions in the study of entrainment: quantifying acoustic-prosodic entrainment at the team rather than the dyad-level, and incorporating the teamwork construct of participation dominance into the study of entrainment. Our current plans include continued corpus development (recall Section 4.4), and using more sophisticated methods than dyad averaging (e.g., using weighting based on team process measures) to move from dyads to teams.

With respect to broader impact, our entrainment measures could be used to mine existing corpora for naturalistic successful and unsuccessful conversations, or to trigger online interventions by dialogue systems participating in multi-party conversations. After additional research understanding the important thresholds for entrainment, organizations could unobtrusively measure team effectiveness during entrainment, and intervene with training to aid teams with low entrainment. Similar interventions would be useful for conversational agents that monitor and facilitate group interactions (e.g., in education via computer-supported collaborative learning). Our work could also support the development of data mining applications for corpora such as team meetings or discussions, from classrooms to boardrooms. Finally, our corpus could support natural language processing research regarding any other aspect of teamwork (e.g., affect, conflict, topic modeling). In sum, the Teams Corpus should provide usable, multi-channel data for examining team processes for a range of purposes and research disciplines.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. 1420784 and 1420377. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors wish to thank Catharine Oertel and Mattias Heldner for advice regarding both lab and equipment needs, and the Learning Research and Development Center at the University of Pittsburgh for lab renovation. Finally, we would like to thank Anish Kumar, the Pitt NLP group, and the anonymous reviewers for their help in improving the paper.

References

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multiparty chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937.

Kelsey Allen, Giuseppe Carenini, and Raymond Ng. 2014. Detecting disagreement in conversations using pseudo-monologic rhetorical structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1169–1180.

Daniel J. Beal, Robin. R. Cohen, Michael J. Burke, and Christy L. McLendon. 2003. Cohesion and performance in groups: A meta-analytic clarification of construct relations. *Journal of Applied Psychology*, 88:989–1004.

Štefan Beňuš, Agustín Gravano, Rivka Levitan, Sarah Ita Levitan, Laura Willson, and Julia Hirschberg. 2014. Entrainment, dominance and alliance in supreme court hearings. *Knowledge-Based Systems*, 71:3–14.

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

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

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The ami meeting corpus: A pre-announcement. In Steve Renals and Samy Bengio, editors, *Machine Learning for Multimodal Interaction*, volume 3869 of *Lecture Notes in Computer Science*, pages 28–39.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, pages 699–708.

James E. Driskell, Eduardo Salas, and Sandra Hughes. 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.

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

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

M. E. Gregory, J. Feitosa, T. Driskell, E. Salas, and W. B. Vessey, 2013. *Developing and enhancing teamwork in organizations: Evidence-based best practices and guidelines*, chapter Designing, delivering, and evaluating team training in organizations. Jossey-Bass, San Francisco.

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

O. P. John, E. M. Donahue, and R. L. Kentle. 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.

Chi-Chun Lee, Athanasios Katsamanis, Matthew P. Black, Brian R. Baucom, Panayiotis G. Georgiou, and Shrikanth Narayanan. 2011. An analysis of pca-based vocal entrainment measures in married couples' affective spoken interactions. In *INTERSPEECH*, pages 3101–3104.

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Interspeech*.

Rivka Levitan, Agustín Gravano, and Julia Hirschberg. 2011. Entrainment in speech preceding backchannels. In *Proceedings of ACL/HLT*, June.

Rivka Levitan, Agustín Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. 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.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2013. Automated two-way entrainment to improve spoken dialog system performance. In *ICASSP*, pages 8372–8376.

Nichola Lubold and Heather Pon-Barry. 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.

Jessica R. Mesmer-Magnus and Leslie A. DeChurch. 2009. Information sharing and team performance: A meta-analysis. *Journal of Applied Psychology*, 94:535–546.

Ella Miron, Miriam Erez, and Eitan Naveh. 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.

Christopher Michael Mitchell, Kristy Elizabeth Boyer, and James C. Lester. 2012. From strangers to partners: Examining convergence within a longitudinal study of task-oriented dialogue. In *SIGDIAL Conference*, pages 94–98.

Seungwhan Moon, Saloni Potdar, and Lara Martin. 2014. Identifying student leaders from mooc discussion forums through language influence. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, pages 15–20.

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

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 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.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational markers of constructive discussions. *arXiv preprint arXiv:1604.07407*.

Catharine Oertel, Kenneth A. Funes Mora, Samira Sheikhi, Jean-Marc Odobez, and Joakim Gustafson. 2014. Who will get the grant?: A multimodal corpus for the analysis of conversational behaviours in group interviews. In *Proceedings of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, pages 27–32.

Susanah B. F. Paletz and Christian D. Schunn. 2011. Assessing group-level participation in fluid teams: Testing a new metric. *Behavioral Research Methods*, 43:522–536.

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

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

Eduardo Salas, Deborah DiazGranados, Cameron Klein, C. Shawn Burke, Kevin C. Stagl, Gerald F. Goodwin, and Stanley M. Halpin. 2008. Does team training improve team performance? a meta-analysis. *Human Factors*, 50:903–933.

Tanmay Sinha and Justine Cassell. 2015. Fine-grained analyses of interpersonal processes and their effect on learning. In *Artificial Intelligence in Education: 17th International Conference*, pages 781–785.

Garold Stasser and William Titus. 1987. Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *Journal of Personality and Social Psychology*, 53:81–93.

Svetlana Stoyanchev and Amanda Stent. 2009. Lexical and syntactic priming and their impact in deployed spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 189–192, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ruth Wageman, J. Richard Hackman, and Erin Lehman. 2005. Team diagnostic survey: Development of an instrument. *Journal of Applied Behavioral Science*, 41:373–398.

Hein Wendt, Martin C. Euwema, and I. J. Hetty van Emmerik. 2009. Leadership and team cohesiveness across cultures. *The Leadership Quarterly*, 20:358–370.