MultiCAT: Multimodal Communication Annotations for Teams

Anonymous ACL submission

Abstract

Successful teamwork requires team members to understand each other and communicate effectively, managing multiple linguistic and paralinguistic tasks at once. Because of the potential for interrelatedness of these tasks, it is important to have the ability to make multiple types of predictions on the same dataset. Here, we introduce Multimodal Communication Annotations for Teams (MultiCAT), a speech- and text-based dataset consisting of audio recordings, automated and hand-corrected transcriptions. MultiCAT builds upon data from teams working collaboratively to save victims in a simulated search and rescue mission, and consists of annotations and benchmark results for the following tasks: (1) dialog act classification, (2) adjacency pair detection, (3) sentiment and emotion recognition, (4) closed-loop communication detection, and (5) phonetic entrainment detection. We also present exploratory analyses on the relationship between our annotations and team outcomes. We posit that additional work on these tasks and their intersection will further improve understanding of team communication and its relation to team performance.

1 Introduction

The last two years have seen an unprecedented rate of advancement in the capabilities of dialog systems. The most recent flagship models from OpenAI (OpenAI, 2024) and Google (Anil et al., 2023) reason across multiple modalities: images, audio, video, and text. Despite these remarkable capabilities, these systems are only capable of 1-on-1 interactions with humans, limiting the potential for their integration into human-machine teams of the future that leverage the complementary strengths of humans and artificially intelligent (AI) agents. Further, these models do not reason about affect, a critical component of team dynamics that is often conveyed via nonverbal information channels, e.g., voice inflection and body language. We assert that next-generation AI systems will require an understanding of *multiparty* dialog (i.e., involving more than two interlocutors), *affect*, and *team dynamics* in order to serve as more effective teammates.

To support the development of these capabilities, we present *Multimodal Communication Annotations for Teams (MultiCAT)*, a novel speech- and text-based dataset that is annotated for sentiment, emotion, dialog acts (DAs), adjacency pairs (APs), phonetic entrainment, and closed-loop communication (CLC) for multiparty dialog in a collaborative search and rescue task. The primary contributions of this paper are the following:

(1) A novel multiparty spoken dialog dataset with annotations for related paralinguistic and conversational classification and regression tasks. To our knowledge, ours is the first publicly available dataset for CLC detection.

(2) Baseline models for detecting entrainment and labeling dialog acts, adjacency pairs, sentiment, emotion, and CLC events. To our knowledge, ours is the first benchmark for unsupervised multi-party entrainment detection.

(3) Exploratory analyses relating our annotations to team outcomes, with results suggesting that our annotations may be better predictors of team performance than participants' self-reported proficiency and expertise.

The rest of the paper is organized as follows. We summarize and motivate the dataset (§ 2). This is followed by sections describing related work, annotation procedures, and benchmark results for individual annotation types (§ 3–§ 6). We then explore the relation between our annotations and team outcomes (§ 7), and conclude in § 8.

2 Dataset

We annotate a subset of the ASIST Study 3 dataset (Huang et al., 2022b,a)—an existing dataset from a large-scale, remotely-conducted human-

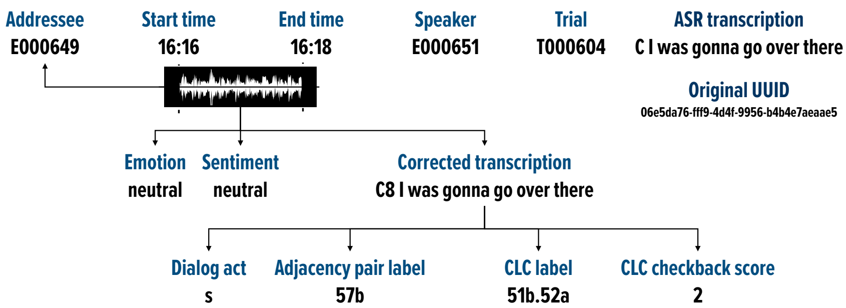


Figure 1: Organization of utterances and labels within the MultiCAT dataset, illustrated by example annotations for a single utterance. The figure also depicts the annotation flow—addressee, emotion, and sentiment annotation and transcript correction are based on the original audio recordings, followed by the corrected transcripts being used for the dialog act, adjacency pair, and CLC annotation tasks. For clarity, we omit IPU annotations in this figure.

machine teaming experiment involving teams of three humans executing simulated urban search-and-rescue (SAR) missions in a Minecraft-based testbed. Each teammate has unique capabilities and information, ensuring that they must communicate with each other to achieve the best results. The goal of the missions is to maximize the team's score, which is based on the number of victims identified, triaged, and moved to a safe zone within a 15-minute time limit.

We chose to annotate this dataset since ASIST Study 3 was designed to elicit teamwork through a combination of complementary roles, capabilities, and knowledge between the three humans on each team. To our knowledge, this dataset is one of only two publicly available datasets in which the dialogs (i) have more than two interlocutors, (ii) are captured using both audio and text (we are interested in both *what* the humans say and *how* they say it, as we believe the latter contains valuable information about social dynamics), (iii) occur in the context of a collaborative team task (we are interested in studying the relation between team communication patterns and team performance), and (iv) is spontaneous and naturalistic (i.e., not using actors, Wizard-of-Oz setups, or synthetic data generation). See Table 18 for a comparison of MultiCAT with a number of related datasets.

Additionally, a Minecraft-based task gives us access to the 'ground-truth' states of the participants (e.g., position, velocity) and their actions (e.g., rescuing a victim). This results in rich behavioral data that can be used to study the interplay between team communication, behavior, and performance. In this paper, we perform an exploratory analysis of the relation between team communication and team performance, but in the future, we plan to perform more fine-grained analyses of team communication, behavior, and performance, and their relationship with each other.

The other dataset that satisfies the aforementioned criteria is the ToMCAT dataset (Pyarelal et al., 2023), which uses the same Minecraft-based SAR task as the ASIST Study 3 dataset, but with in-person participants instrumented with physiological sensors, rather than remote participants.

We annotate a subset of the ASIST Study 3 dataset for sentiment, emotion, dialog acts, adjacency pairs, closed-loop communication events, utterance addressee, and interpausal unit boundaries (see Figure 1). In addition, we provide corrected gold transcriptions for the conversations, which originally had ASR-generated transcriptions.

Data collection procedure Participants are recruited from a pool of adults in the US who play Minecraft and speak English. Selected participant demographic details are provided in Table 8. Participants fill out a series of surveys related to their background with Minecraft, their leadership style, and sociological factors that may impact their performance in the study. They then carry out a training mission, followed by two separate missions with the same team, either on their own or with a human or AI advisor assisting them. The two missions differ in the layout of the environment and the location of the victims to be rescued.

Participants use their own computer for the task, and as such their setups may vary. Their speech is recorded on separate channels, with utterance-level transcriptions obtained in real time using Google's enhanced phone call speech to text model.[1] Participants were compensated with either a $35 Amazon

[1]`https://cloud.google.com/speech-to-text/docs/enhanced-models`

gift card or course credit. If they were unable to complete the study due to technological issues, they were compensated at the rate of $15 per hour, rounded up to the nearest hour.

Annotation procedure The starting point for data in MultiCAT is a set of utterance-aligned speech and text transcriptions. We trained five annotators who completed annotation tasks that matched their expertise (see § B.4 for details). The annotators were all native or highly proficient English speakers, and were paid the standard hourly student wage set by their respective universities. They underwent an iterative training procedure while working to achieve task-specific acceptable levels of agreement on a small portion of the data (the annotations from the training phase are not included in the dataset); subsequent annotations were completed by one annotator each.

Dataset overview The dataset is structured as follows. All utterances have a unique identifier (UUID) generated as part of the ASR transcription process, with the exception of a relatively small number of utterances (401) that were inserted as part of the manual transcript correction process—these can nevertheless be uniquely identified by combining their trial ID, participant ID, and start timestamp. Each item is associated with its speaker, the mission in which it was created, and the start and end times of the utterance. Along with the task-specific labels, we also annotate instances of background noises.

A closer examination of the dataset (see Table 1 for details) reveals its particular benefits for the end user. The dataset contains a total of 11,024 utterances. Trials vary in amount of communication, ranging from 91 to 348 utterances. There is further variability in the amount of conversation attributed to an individual team member, with the number of utterances ranging from 19 to 156. This variability lends itself to an exploration of the dynamics of teamwork, different types of team members, and their relationships with team performance.

Differing numbers of trials were used for annotating different tasks due to small minority classes (emotion and sentiment annotation) and the difficulty of annotation (IPU boundary and addressee annotation). A detailed breakdown of which trials are annotated for which tasks can be found in Appendix D. The total numbers of items in MultiCAT with each label for each task are provided in Appendix C.

The MultiCAT dataset is included in the supplementary material in the form of an SQLite3 database (`multicat.db`). Along with the annotations, the database contains the following data from the original ASIST Study 3 dataset in order to facilitate analyses: the original ASR utterance transcriptions and their UUIDs, participant demographic details, and participants' self-reported gaming proficiency and experience, the final team score, and the advisor assigned to the team. We do not include the original audio files in the MultiCAT dataset—they can be obtained from the ASIST Study 3 dataset.

The MultiCAT dataset is licensed under the Creative Commons 4.0 BY license (CC BY 4.0).

3 Dialog acts and adjacency pairs

Related work A dialog act (DA) is the communicative function underlying a speaker's utterance (Bunt et al., 2020). While numerous annotated resources are available for DAs, their annotation schemes vary depending on their purpose, such as capturing domain-specific phenomena. The Switchboard Dialog Act (SwDA) (Jurafsky et al., 1997) and the Meeting Recorder Dialog Act (MRDA) (Shriberg et al., 2004) corpora are both based on naturally occurring conversations, and use the DAMSL (Core and Allen, 1997) tag-set with some modifications—an approach we adopt as well. While the SwDA corpus contains dyadic dialog, the MRDA dataset contains multi-party (defined as involving more than two interlocutors) dialog.

DailyDialog (Li et al., 2017) is a text-based dataset using short human-written dyadic dialogs that follows Amanova et al. (2016). This dataset differs from ours in two notable ways. First, while DailyDialog contains annotations for only four DA labels, we use many more DA labels since we are interested in more fine-grained intentions. Second, the conversations in the DailyDialog corpus are more formal and less task-oriented compared to the conversations in our dataset that are naturalistic and occur in the context of a collaborative task. The STAC corpus (Asher et al., 2016) annotations capture the dialog structure in a multiparty setting. The communication occurs over a chat interface where the participants play a non-cooperative game with opposing goals. We capture the conversation flow by means of adjacency pairs.

Annotation procedure For our annotations of dialog acts (DAs), we used the framework from the MRDA dataset, which, like MultiCAT, consists of

Quantity	Total
Trials	49
Teams	25
Speakers	73
Utterances	11024
Word types	2607
Word tokens	108475

(a) Totals

	Mean	Min	Max	SD
Utts./spkr	151	42	287	54
Utts./trial	225	91	348	65
Utts./spkr/trial	79	19	156	28
Word types/utt.	9	1	74	8
Word tokens/utt.	10	1	118	11

(b) Mean, minimum, maximum, and SD.

Annotation	# Trials	# Utts
Emotion	46	7731
Sentiment	46	7731
CLC	36	6544
Gold transcript	45	4666
Dialog act	45	10342
APs	45	6846
Entrainment	8	2896

(c) Number of trials and annotated utterances for our annotation types.

Table 1: Highlights of the MultiCAT dataset. Not all utterances receive labels for all the tasks. AP, DA, and CLC tasks; only items with valid labels are counted here.

natural task-oriented human conversations. Under this framework, each utterance is annotated with a 'general' and zero or more 'specific' tags. Due to imperfect segmentation by the ASR system, our data contained single utterances that should have been split up into multiple utterances. To align the DA annotations with the rest of the annotation tasks while still letting an utterance have more than one DA label, we use the pipe symbol (|) to indicate segmentation. Finally, since inter-annotator agreement on the Accept (aa) and Acknowledgment (bk) tags was very low, we merged them into a single tag (aa). In total, there are 11 general tags and 38 specific tags[2]. The inter-annotator agreement measured using Cohen's κ is 0.6238 for the general DA category.

Adjacency pairs We also annotate the conversational structure in the dialogs using the conventions for adjacency pairs (APs) presented in MRDA (Dhillon et al., 2004). APs capture paired utterances such as question-answer, greeting-greeting, etc. An AP for a sequence of utterances is defined such that it contains two parts, each containing one or more utterances and uttered by different speakers (Levinson et al., 1983).

Baseline models We provide two baseline model results: He et al.'s (2021) and LLaMA-3[3]. We include results for the 50 fine-grained and 5 coarse-grained labels[4] on the corrected transcripts. Since this is a highly imbalanced dataset, we report the macro F1 score along with the accuracy in Table 2. For the LLaMA-3 baseline, we report the mean of three random runs, with the standard error of the mean (SEM) in parentheses. See Appendix G for further details on the model training.

[2]We do not annotate for rising tone (rt), which is a non-DA tag.

[3]`https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`

[4]The 5 coarse-grained tags are Statement, Filler, Backchannel, Disruption, and Question.

4 Closed-loop communication

Related work Good teamwork processes enable teams to perform beyond the sum of their parts (Roberts et al., 2021). Closed-loop communication (CLC) has been proposed in the team science literature as one of the coordinating mechanisms for effective teamwork (Salas et al., 2005). This communication strategy has been implemented in military contexts to reduce the frequency of communication breakdowns in teams (Burke et al., 2004), and is being explored in the context of healthcare as well (Parush et al., 2011). CLC has been shown to be correlated with improved outcomes in both simulations (Diaz and Dawson, 2020) and the real world (Härgestam et al., 2013; El-Shafy et al., 2018), with studies suggesting that high-performing teams tend to display CLC more often than low-performing teams (Bowers et al., 1998), and that deviations from CLC can lead to information loss (Parush et al., 2011) and degraded task performance (Lieber et al., 2022). These findings suggest the utility of developing methods to automatically detect deviations from CLC protocols in real-time, in order to provide appropriate interventions—e.g., an AI agent that informs the team in a timely manner when there is a communication breakdown.

Automated CLC detection is a relatively understudied task. Rosser et al. (2019) developed an NLP-based method to identify CLC and found positive relationships between the outputs of their algorithm and annotations performed by a trained human annotator. However, we were not able to find further details on their method or dataset. Winner et al. (2022) assess the usability of a 'Team Dynamics Measurement System' (TDMS) proto-

	Fine-grained		Coarse-grained	
Model	Macro F1 (SEM)	Accuracy (%)	Macro F1 (SEM)	Accuracy
He et al.'s (2021)	30.75	63.24	42.15	93.92
LLaMA-3	34.76 (0.48)	66.47 (0.15)	44.55(0.90)	94.66 (0.07)

Table 2: Macro F1 and accuracy for DA classification on fine-grained and coarse-grained classes.

type, which implements a measure of CLC that relies solely on communication flow data (e.g., interlocutor identity, utterance timing, and turn-taking patterns), while ignoring the actual content of the utterances. Robinson et al. (2023b) improve upon the flow-based measure by incorporating keyword analysis to analyze the content of the utterances. The dataset used for both of these studies (Robinson et al., 2023a) is not publicly available, limiting our ability to compare our work to theirs.

Though varying definitions of CLC can be found in the literature (Diaz and Dawson, 2020; Salik and Ashurst, 2022; Salas et al., 2005; Marzuki et al., 2019; Härgestam et al., 2013), most definitions of what we refer to as a CLC 'event' include the following three sub-events occurring in sequence: (1) *Call-out*: Interlocutor I_1 shares information with/gives an instruction to interlocutor I_2 (Butcher, 2018), (2) *Check-back*: I_2 confirms their understanding of the information/instruction by repeating it back to I_1, and (3) *Closing*: I_1 confirms that I_2 has received and understood the information or performed the desired action.

To our knowledge, MultiCAT is the first publicly available dataset for studying CLC. Most existing CLC research is conducted by watching videos and recording only the parts that researchers are interested in (e.g., CLC categories (Marzuki et al., 2019) and task completion time (El-Shafy et al., 2018)) without annotating the entire dialog.

Annotation procedure Annotators were trained to identify and label CLC sub-events and score the quality of check-backs on a scale of 1–3, as detailed in Table 26. We used `a`, `b`, and `c` to denote call-outs, check-backs, and closings, respectively, to partially align our CLC labels with the labels for AP components. The inter-annotator agreement calculated using Krippendorff's α was 0.68, which we deemed acceptable given the challenging nature of this annotation task, which involves a nontrivial amount of subjective interpretation, dealing with ambiguity, and keeping large windows of utterances in the annotator's working memory.

Baseline Model We use a three-stage approach to identifying CLC events.

In the first step, we construct TF-IDF feature vectors from lemmatized versions of the utterances, which are then used as inputs to a logistic regression model that predicts whether or not an utterance corresponds to a call-out sub-event (i.e., `a`). Second, for each utterance that is labeled as a call-out, we examine the next three utterances following that utterance that are from a speaker other than the source of the call-out utterance. For each of the call-outs and their three candidate check-back pairs, we use a RoBERTa-based sequence classification model fine-tuned on MultiCAT to predict whether the candidate utterances check back to the call-out utterance (i.e., `b`).

Third, given the rarity of 'closing' sub-events, we combine subevent sequences `ab` and `abc` into a single CLC event category, contrasting it against isolated call-outs classified as 'open loop events'. This pragmatic categorization is consistent with the prevalence of two-stage CLC events in real-world scenarios noted by Robinson et al. (2023b) and Marzuki et al. (2019).

We aggregated the labels from the previous steps to classify the overall CLC event status into three categories: *closed-loop event*, *open-loop event*, and *non-CLC event*. For every utterance, if a call-out sub-event is detected, and if at least one check-back is detected among the next three utterances from speakers other than the original speaker, we conclude that this call-out is 'closed' and a CLC event has occurred. Conversely, if no check-back is detected then the call-out by itself forms an open-loop event. Non-CLC events are categorized as situations where the initial call-out is not detected at all. Results for all three stages are provided in Table 3, and details on our model training are provided in Appendix G.

5 Sentiment/Emotion recognition

Related work Datasets for sentiment and emotion have largely been annotated for one or both tasks, but not others. GEMEP (Bänzinger et al., 2012) and

Stage	Accuracy	F_1
Call-out detection	.77	.79
Check-back detection	.76	.43
Complete CLC event detection	.51	.45

Table 3: Results for the CLC detection baseline approach. For the complete CLC event detection stage, we report a weighted F_1 score due to the very small number of 'closing' sub-events in the data.

		Models		
Sentiment	Support	Strat.	Multi.	LLaMA-3
Negative	370	15.0	43.5	52.1 (0.0)
Neutral	1310	51.5	62.7	68.8 (0.0)
Positive	611	28.0	49.8	54.0 (0.0)
All	2291	31.5	52.0	58.4 (0.24)

Table 4: Results for sentiment prediction.

IEMOCAP (Busso et al., 2008) contain a total of 10 actors each simulating a range of emotions. Both contain high-quality recordings but are relatively small corpora. RAVDESS (Livingstone and Russo, 2018) likewise contains actors simulating emotion, with an additional annotation for the intensity of the emotion. The YouTube dataset (Morency et al., 2011) contains 47 videos of single speakers, with utterances annotated for sentiment. Similarly, ICT-MMMO (Wöllmer et al., 2013) contains single-speaker data annotated for sentiment, with each item being relatively long.

The Multimodal Emotion Lines Dataset (MELD) (Poria et al., 2019) consists of conversations from the TV show *Friends* and is annotated for Ekman's universal emotions (Ekman, 1992) and positive, negative, or neutral sentiment. Likewise, the CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset (Bagher Zadeh et al., 2018) is annotated for both tasks, with seven sentiment labels ranging from strong negative to strong positive. CMU-MOSEI uses monologue data from YouTube. DailyDialog is also annotated for Ekman's universal emotions. While all of these datasets contain annotation types that have some overlap with those present in MultiCAT, none contain the range we present here.

Annotation procedure Two annotators were trained to identify the opinions of the speaker towards the subject (sentiment) and the affect shown by the speaker (emotion) during an utterance, by listening to it in context. Inter-annotator agreement was calculated using Cohen's κ; annotators achieved an agreement score of 0.89 for sentiment and 0.83 for emotion. We use the same set of emotions as MELD and DailyDialog, namely Ekman's universal emotions (Ekman, 1992)—anger, disgust, fear, joy, sadness, and surprise—along with neutral. Sentiment labels are positive, negative, and neutral.

Baseline models We provide results for three baseline models. The first, 'Stratified', is a classifier that predicts classes with probabilities proportional to their proportion in the training set.

The second, 'Multitask', is a multitask sentiment and emotion classifier based on Culnan et al.'s (2021) model, which uses low-level acoustic features from the Interspeech 13 feature set created for tasks including emotion and social cues (Schuller et al., 2013) extracted with openSMILE (Eyben et al., 2010). We use 768-d word embeddings generated with BERT (Devlin et al., 2019) model bert-base-uncased as text features. Text is fed through a bidirectional LSTM, while acoustic features are averaged and fed through feedforward layers. The output of these two components are then concatenated and fed through two feedforward layers to reduce their dimension to 100. Finally, the output of these two layers is passed to task-specific heads to make sentiment and emotion predictions.

The model is pretrained on data from MELD and CMU-MOSI. CMU-MOSI contains sentiment labels from strong negative to strong positive, so we collapse over negative and positive label types to get the same three classes of interest as in MultiCAT. We also provide a third baseline, based on LLaMA-3, but only using text as input, without audio. Results for sentiment and emotion tasks are provided in Table 4 and Table 5.

We report F1 for each class and overall macro F1 for all classes. For the LLaMA-3 baseline, the results are based on the mean of three random runs, with the SEM in parentheses. We also provide the number of items per class and the overall number of items in the 'Support' column. More details can be found in Appendix G.

We find that our multitask sentiment and emotion prediction model is more successful at predicting sentiment than it is at predicting emotion, with better performance for majority classes than minority classes. In the case of emotion prediction, difficulty arises from two very small minority classes: anger and disgust.

		Models		
Emotion	Support	Strat.	Multi.	LLaMA-3
Anger	18	5.4	0.0	3.9 (0.03)
Disgust	25	0.0	9.3	15.8 (0.0)
Fear	70	3.2	16.2	27.2 (0.02)
Joy	154	4.2	20.1	19.6 (0.01)
Neutral	1799	77.5	76.5	87.7 (0.0)
Sadness	145	5.6	30.5	36.7 (0.01)
Surprise	80	3.7	29.2	31.6 (0.02)
All	2291	14.2	26.0	31.8 (0.92)

Table 5: Results for emotion prediction.

6 Entrainment detection

Entrainment is the adaption of verbal and non-verbal actions by conversation partners to more closely resemble one another (Borrie and Liss, 2014). It facilitates effective turn taking, builds rapport, and aids in communicating positive sentiments. Correlations between entrainment and desired social outcomes have been reported in cooperative games (Yu et al., 2019; Levitan et al., 2015), patient-therapist relations (Nasir et al., 2020; Borrie et al., 2019), study groups (Friedberg et al., 2012), and romantic success (Ireland et al., 2011). Besides English, entrainment has been studied in Hebrew (Weise et al., 2022), Russian (Kachkovskaia et al., 2020; Menshikova et al., 2020), Slovak, Spanish, and Chinese (Levitan et al., 2015) as well.

The study of entrainment faces many challenges. Many popular corpora have relatively a modest number of teams—e.g., the Columbia Games Corpus[5] and the Brooklyn Multi-Interaction Corpus (Weise et al., 2022) have 12 each (compared to the 49 teams in MultiCAT). Some are also restricted due to being sensitive in nature, e.g., the Suicide Risk Assessment Corpus (Baucom et al., 2014) and the Couples Therapy Corpus (Christensen et al., 2004), or prohibitively expensive to obtain, e.g., the Fisher Corpus (Cieri et al., 2004).

Prior work has relied on pristine recording conditions with professional recording equipment and manual preparation of an acoustic-prosodic feature set, restricting entrainment-specific datasets to laboratory conditions. In contrast, MultiCAT is based on data collected in more realistic conditions, where researchers exert limited control over recording channels, environments, and participant interactions. MultiCAT also enables the analysis of entrainment in short-lived, randomly formed teams in which the teammates do not know each other beforehand.

Annotation procedure Previous research on vocalic entrainment has concentrated on dyadic interactions with balanced turn-taking and responses directed at one intended listener. However, the distribution of utterances in a multi-party conversation is less likely to be balanced than in a dyadic conversation. Additionally, in a multi-party conversation, utterances could be aimed at the group as a whole, rather than one intended listener. Thus, there is a need to identify speaker dyads and separate them from utterances with no specific intended listener.

We identify the subset of utterances in three-member trials in which there is a single intended addressee to find dyadic interactions within a multi-party conversation. For this annotation task, 4 teams (8 trials), were randomly selected. Annotators completed this annotation task in Praat (Boersma, 2001), using each speaker's individual audio stream (in order to avoid speaker overlap), gold transcriptions, and Praat textgrids. One of the eight selected trials (T000605) is missing audio data for one speaker, thus yielding data 11 unique speakers.

For each trial, annotators identified the boundaries of a stream of audio separated by a pause of 50ms or more, also known as an inter-pausal unit (IPUs). Next, they mapped the audio in each IPU to the corresponding text from the transcript (an utterance can have one or more IPUs), and identified the addressee of each IPU. The addressee labels had four possibilities—an identifier for each of the three participants, or 'all' to indicate a general response or an unknown audience. Annotators achieved a Cohen's κ score of 0.48.

Baseline We replicate the baseline model used in Nasir et al. (2020) for assessment of their unsupervised model, using the same training corpus, acoustic feature set and hyperparameters. First, 80% of the utterances from the Fisher Corpus English Part 1 (LDC2004S13) (Cieri et al., 2004) are randomly chosen. An encoder-decoder model is used to encode entrainable information from a given utterance and predict the next turn, which is compared to its referent (i.e., the real 'next turn') to compute the loss.

To verify if this model is able to detect entrainment in a multi-party system, we use the verification measures from Nasir et al. (2020), in which the model classifies conversations as 'real' (all pairs of adjacent utterances are in order) or 'fake' (turns

[5] http://www.cs.columbia.edu/speech/games-corpus/

scrambled so that the entrainment information is not preserved) when presented with sample conversations from the test set. First, dyadic interactions are extracted using the addressee labels for each of the 8 trials (8 × 3 = 24 possibilities). This yields 11 interactions, a number lower than the expected number (23) since not all participants were judged to have addressed both their team mates. Turn-level acoustic features are then extracted and processed to function as a test set for the model.

The classification accuracy for the MultiCAT entrainment set was 51.86% (mean of 30 runs). This is much lower than the accuracies achieved by Nasir et al. (2020) for the two-party Fisher test set and Suicide Corpus (72.10% and 70.44% respectively). This may be due to two factors. First, the increase in the number of interlocutors from two to three increases the complexity of detecting entrainment. Second, the differences in the recording conditions for the training corpus and the MultiCAT corpus (controlled vs real-world) pose a challenge to detecting vocalic entrainment, an effect that is sensitive to recording conditions. Despite the lower accuracy, we choose to report these results because to the best of our knowledge, there are no existing benchmarks for unsupervised *multi-party* entrainment detection.

7 Annotations and team outcomes

We examined the relationship between our annotations and team outcomes by developing baseline models for predicting the final team score at the end of a mission.

For each trial, we constructed eight sets of features—(i) five containing the counts of different label types ('AP', 'CLC', 'DA', 'Sentiment', and 'Emotion') for utterances in that trial, (ii) the union of these five sets ('MultiCAT'), (iii) a set of features constructed from participants' self-reported proficiency and expertise ('Proficiency'), and (iv) the union of the seven aforementioned sets ('All'). Further details are provided in Appendix F. Features are scaled to zero mean and unit variance. We then perform principal components analysis (PCA) and use the component with the highest variance as a predictor for ridge regression models (see § G.6 for details).

Table 6 shows results for our score prediction models using the eight feature sets described earlier. We evaluate our models using leave-one-out cross-validation (LOOCV) and report the mean absolute error (MAE) across all folds as well as the SEM.

	Mission 1	Mission 2	Combined
# of trials	17	16	33
Proficiency	130 (26)	104 (19)	118 (17)
AP	126 (17)	99 (13)	117 (12)
CLC	124 (13)	97 (10)	117 (9)
DA	124 (11)	99 (9)	117 (8)
Sentiment	124 (10)	100 (8)	116 (7)
Emotion	123 (9)	98 (7)	115 (6)
MultiCAT	123 (8)	97 (7)	115 (6)
All	123 (8)	97 (6)	115 (5)

Table 6: MAE (with SEM in parentheses) over all LOOCV folds for our score prediction models.

For this analysis, we restrict ourselves to trials that contain DA, AP, CLC, sentiment, and emotion labels.

The MAE for feature sets that include our annotations is lower than that for the Proficiency feature set, suggesting that our annotations may be better predictors of team performance than self-reported proficiency and experience. We do not make a strong claim here though, since the error bars (± SEM) overlap. Note, however, that the error bars for the Proficiency set are consistently larger than the error bars for models including our annotations as features. Combining the Proficiency and MultiCAT sets does not reduce the MAE, but it does reduce the SEM for the Mission 2 and Combined trial sets.

We also find that the MAEs for Mission 2 are better than those for Mission 1. This may be due to the participants still getting used to the task and their teammates in the first mission, thereby suppressing the effects of differences in proficiency and team communication. This is consistent with the results of Soares et al. (2024), who found that their model of interpersonal coordination was more predictive of team performance in Mission 2 compared to Mission 1. Notably, their model uses semantic and vocalic features from team dialog, and was evaluated on both the ASIST Study 3 and ToMCAT datasets, further supporting the connection between multimodal team dialog and team performance.

8 Conclusion

We present MultiCAT, a dataset annotated for six computational tasks that may be studied individually or in concert to make assessments about team outcomes. We also demonstrate MultiCAT's usefulness for tasks involving individual annotation types as well as downstream tasks involving multiple annotation types, and provide baseline models for comparison with future research.

9 Limitations

As with any novel dataset, MultiCAT has its limitations. First, data is only in English, largely from native speakers of American English. Conclusions drawn from and patterns found in this dataset may not generalize to other languages or populations.

Additionally, because natural language does not have an equal distribution of items from all dialog act classes, for example, and because each emotion does not appear with equal frequency, datasets consisting of conversations of unconstrained natural language that are created for these tasks will be inherently imbalanced. This is true of MultiCAT, as well. This limitation necessarily affects models seeking to make good predictions about minority classes, as there may be few examples of a given minority class.

Finally, the score prediction models in § 7 are fairly basic ridge regression models. While this can be a strength in terms of interpretability, it is possible that more sophisticated models can better capture the relationship between our annotations and team performance.

We believe that acknowledging these limitations in future research will help avoid the risks of overgeneralizing results to other populations and making assumptions about patterns of data in non-English languages.

10 Ethics Statement

In this work, we annotated a subset of the publicly available ASIST Study 3 dataset (Huang et al., 2022b). Our use of the dataset is consistent with its terms of use (CC0 1.0).

Both the collection of the ASIST Study 3 dataset and our analysis of it were approved by IRBs. Participants in the ASIST Study 3 dataset were voluntary participants who signed informed consent forms and were aware of any risks of harm associated with their participation.

The dataset collection process and conditions are described in § 2. The group of annotators was comprised of three graduate students and one undergraduate student. All annotators were compensated fairly for their time in accordance with the standard hourly wages set by their respective departments (in the case of graduate students) or their university (in the case of the undergraduate student).

The characteristics of the dataset are provided in Appendix B. We provide information about the compute resources required for model training in Appendix G.

Intended use If our technology functions as intended, it could be deployed as part of social AI agents embedded in human-machine teams—these agents would be able to understand the affective states of their human teammates, as well as social dynamics within the team.

Failure modes Failure modes of our technology involve incorrect predictions. It is conceivable (in the context of human-machine teaming) that deteriorated outcomes may result from ineffective human-machine teaming that occurs due to a social AI agent's inability to understand their human teammates.

Misuse potential It is also conceivable that malicious actors may endow AI agents with the ability to infer sentiment, emotion, team dynamics, etc. in order to perform social engineering for nefarious purposes.

Collecting data from users We are not proposing a system to collect data from users in this paper.

Potential harm to vulnerable populations To our knowledge, the possible harms we have identified are not likely to fall disproportionately on populations that already experience marginalization or or otherwise vulnerable.

References

Dilafruz Amanova, Volha Petukhova, and Dietrich Klakow. 2016. Creating annotated dialogue resources: Cross-domain dialogue act classification. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 111–117, Portorož, Slovenia. European Language Resources Association (ELRA).

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Tanja Bänzinger, Marcello Mortarillo, and Klaus R Scherer. 2012. Introducing the Geneva Multimodal Expression Corpus for Experimental Research on Emotion Perception. *Emotion (Washington, D.C.)*, 12(5):1161–1179.

BR Baucom, AO Crenshaw, CJ Bryan, TA Clemans, TO Bruce, and MD Rudd. 2014. Patient and clinician vocally encoded emotional arousal as predictors of response to brief interventions for suicidality. *Brief Cognitive Behavioral Interventions to Reduce Suicide Attempts in Military Personnel. Association for Behavioral and Cognitive Therapies*.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Stephanie A Borrie, Tyson S Barrett, Megan M Willi, and Visar Berisha. 2019. Syncing up for a good conversation: A clinically meaningful methodology for capturing conversational entrainment in the speech domain. *Journal of Speech, Language, and Hearing Research*, 62(2):283–296.

Stephanie A Borrie and Julie M Liss. 2014. Rhythm as a coordinating device: Entrainment with disordered speech. *Journal of Speech, Language, and Hearing Research*, 57(3):815–824.

Clint A. Bowers, Florian Jentsch, Eduardo Salas, and Curt C. Braun. 1998. Analyzing communication sequences for team training needs assessment. *Human Factors*, 40:672+. Article.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation, second edition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 549–558, Marseille, France. European Language Resources Association.

C S Burke, Eduardo Salas, K Wilson-Donnelly, and H Priest. 2004. How to turn a team of experts into an expert medical team: guidance from the aviation and military communities. *BMJ Quality & Safety*, 13(suppl 1):i96–i104.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Brad W Butcher. 2018. Leadership and Crisis Management. *Rapid Response System: A Practical Guide*, 19.

Andrew Christensen, David C Atkins, Sara Berns, Jennifer Wheeler, Donald H Baucom, and Lorelei E Simpson. 2004. Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. *Journal of consulting and clinical psychology*, 72(2):176.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher corpus: A resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

Mark G Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, volume 56, pages 28–35. Boston, MA.

John Culnan, Seongjin Park, Meghavarshini Krishnaswamy, and Rebecca Sharp. 2021. Me, myself, and ire: Effects of automatic transcription quality on emotion, sarcasm, and personality detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 250–256, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. 2004. Meeting recorder project: Dialog act labeling guide. Technical report, International Computer Science Institute, Berkley CA.

Maria Carmen G. Diaz and Kimberly Dawson. 2020. Impact of simulation-based closed-loop communication training on medical errors in a pediatric emergency department. *American journal of medical quality*, 35(6):474–478.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Ibrahim Abd El-Shafy, Jennifer Delgado, Meredith Akerman, Francesca Bullaro, Nathan A. M. Christopherson, and Jose M. Prince. 2018. Closed-loop communication improves task completion in pediatric trauma resuscitation. *Journal of surgical education*, 75(1):58–64.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Heather Friedberg, Diane Litman, and Susannah BF Paletz. 2012. Lexical entrainment and success in student engineering groups. In *2012 IEEE spoken language technology workshop (SLT)*, pages 404–409. IEEE.

Maria Härgestam, Marie Lindkvist, Christine Brulin, Maritha Jacobsson, and Magnus Hultin. 2013. Communication in interdisciplinary teams: exploring closed-loop communication during in situ trauma team training. *BMJ open*, 3(10):e003525.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Zihao He, Leili Tavabi, Kristina Lerman, and Mohammad Soleymani. 2021. Speaker turn modeling for dialogue act classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2150–2157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Heeman and James Allen. 1995. The TRAINS 93 dialogues.

Lixiao Huang, Jared Freeman, Nancy Cooke, John Colonna-Romano, Matthew D Wood, Verica Buchanan, and Stephen J Caufman. 2022a. Exercises for Artificial Social Intelligence in Minecraft Search and Rescue for Teams.

Lixiao Huang, Jared Freeman, Nancy Cooke, John "JCR" Colonna-Romano, Matt Wood, Verica Buchanan, and Stephen Caufman. 2022b. Artificial Social Intelligence for Successful Teams (ASIST) Study 3.

Molly E Ireland, Richard B Slatcher, Paul W Eastwick, Lauren E Scissors, Eli J Finkel, and James W Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological science*, 22(1):39–44.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13. Technical report, University of Colorado at Boulder and +SRI International.

Tatiana Kachkovskaia, Tatiana Chukaeva, Vera Evdokimova, Pavel Kholiavin, Natalia Kriakina, Daniil Kocharov, Anna Mamushina, Alla Menshikova, and Svetlana Zimina. 2020. SibLing corpus of Russian dialogue speech designed for research on speech entrainment. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6556–6561, Marseille, France. European Language Resources Association.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2023. Delidata: A dataset for deliberation in multi-party problem solving. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen C Levinson, Stephen C Levinson, and S Levinson. 1983. *Pragmatics*. Cambridge University Press.

Rivka Levitan, Štefan Beňuš, Agustín Gravano, and Julia Hirschberg. 2015. Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross-linguistic comparison. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 325–334.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Christopher S. Lieber, Yancy Vance Paredes, Aaron Zhen Yang Teo, and Nancy J. Cooke. 2022. Analysis of Voice Transmissions of Air Traffic Controllers in the Context of Closed Loop Communication Deviation and its Relationship to Loss of Separation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 66(1):672–676.

Steven R. Livingstone and Frank A. Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391–e0196391.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Ernisa Marzuki, Hannah Rohde, Chris Cummins, Holly Branigan, Gareth Clegg, Anna Crawford, and Lisa MacInnes. 2019. Closed-loop communication during out-of-hospital resuscitation: Are the loops really closed? *Communication and Medicine*, 16(1):54–66.

Alla Menshikova, Daniil Kocharov, and Tatiana Kachkovskaia. 2020. Phonetic Entrainment in Cooperative Dialogues: A Case of Russian. In *Proceedings of Interspeech 2020*, pages 4148–4152.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176.

Md Nasir, Brian Baucom, Craig Bryan, Shrikanth Narayanan, and Panayiotis Georgiou. 2020. Modeling vocal entrainment in conversational speech using deep unsupervised learning. *IEEE Transactions on Affective Computing*.

OpenAI. 2024. GPT-4o.

Avi Parush, Chelsea Kramer, Tara Foster-Hunt, Kathryn Momtahan, Aren Hunter, and Benjamin Sohmer. 2011. Communication and team situation awareness in the or: Implications for augmentative information display. *Journal of Biomedical Informatics*, 44(3):477–485. Biomedical Complexity and Error.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Volha Petukhova, Martin Gropp, Dietrich Klakow, Gregor Eigner, Mario Topf, Stefan Srb, Petr Motlícek, Blaise Potard, John Dines, Olivier Deroo, Ronny Egeler, Uwe Meinz, Steffen Liersch, and Anna Schmidt. 2014. The DBOX Corpus Collection of Spoken Human-Human and Human-Machine Dialogues. In *International Conference on Language Resources and Evaluation*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536. Association for Computational Linguistics.

Adarsh Pyarelal, Eric Duong, Caleb Shibu, Paulo Soares, Savannah Boyd, Payal Khosla, Valeria A. Pfeifer, Diheng Zhang, Eric Andrews, Rick Champlin, Vincent Raymond, Meghavarshini Krishnaswamy, Clayton T. Morrison, Emily Butler, and Kobus Barnard. 2023. The tomcat dataset. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Aaron P. J. Roberts, Leonie V. Webster, Paul M. Salmon, Rhona Flin, Eduardo Salas, Nancy J. Cooke, Gemma J. M. Read, and Neville A. Stanton. 2021. State of science: Models and methods for understanding and enhancing teams and teamwork in complex sociotechnical systems. *Ergonomics*, pages 1–45.

F. Eric Robinson, Lt Col Sarah Huffman, Lt Col Daniel Bevington, DeAnne French, Clayton Rothwell, LTC Christopher Stucky, Marissa Tharp, and Ashton Hughies. 2023a. Team coordination style is an adaptive, emergent property of interactions between critical care air transport team personnel. *Air medical journal*, 42(3):174–183. ObjectType-Article-1.

Frank E Robinson, David Grimm, Dain Horning, Jamie C Gorman, Jennifer Winner, and Christopher Wiese. 2023b. Using Natural Language Processing to Develop an Automated Measure of Closed Loop Communication Among Critical Care Air Transport Teams.

Alexandra Rosser, Sarah Sullivan, Ryan Thompson, and Hee Soo Jung. 2019. 1774: Automated natural language processing of closed-loop communication in trauma resuscitations. *Critical care medicine*, 47(1 Suppl 1):860–860.

Eduardo Salas, Dana E. Sims, and C. Shawn Burke. 2005. Is there a "big five" in teamwork? *Small Group Research*, 36(5):555–599.

Irim Salik and John V. Ashurst. 2022. *Closed Loop Communication Training in Medical Simulation*. StatPearls Publishing, Treasure Island (FL).

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Elizabeth Shriberg, Rajdip Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the SIGDIAL 2004 Workshop, The 5th Annual Meeting of the Special Interest Group on Discourse and Dialogue, April 30- May 1, 2004, Cambridge, Massachusetts, USA*, pages 97–100. The Association for Computer Linguistics.

Paulo Soares, Adarsh Pyarelal, Meghavarshini Krishnaswamy, Emily Butler, and Kobus Barnard. 2024. Probabilistic modeling of interpersonal coordination processes. In *Forty-first International Conference on Machine Learning*.

Andreas Weise, Matthew McNeill, and Rivka Levitan. 2022. The Brooklyn Multi-Interaction Corpus for Analyzing Variation in Entrainment Behavior. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1721–1731, Marseille, France. European Language Resources Association.

Jennifer Winner, Jayde King, Jamie Gorman, and David Grimm. 2022. Team coordination dynamics measurement in enroute care training: Defining requirements and usability study. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, 11(1):21–25.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

Mingzhi Yu, Diane Litman, and Susannah Paletz. 2019. Investigating the relationship between multi-party linguistic entrainment, team characteristics and the perception of team social outcomes. In *The Thirty-Second International Flairs Conference*.

A Introduction

In these appendices we provide additional details on the dataset (Appendix B, Appendix C, Appendix D), comparison to other datasets (Appendix E), feature engineering (Appendix F), model training (Appendix G), annotation procedures (Appendix I, Appendix J, Appendix K, Appendix L, Appendix M),

B Data Statement

B.1 Curation Rationale

The ASIST Study 3 dataset contains data from eight experimental conditions: (i) teams with no advisor, (ii) teams with human advisors, and (iii) teams with one of six AI advisors (i.e., six conditions). Of these, we opted to exclude trials with human advisors for two reasons: (i) unlike with the actual study participants, we did not have source-separated audio streams for the human advisors, who were

Advisor	# of Trials
None	31
ASI-CMURI-TA1	2
ASI-CRA-TA1	2
ASI-DOLL-TA1	2
ASI-SIFT-TA1	2
ASI-UAZ-TA1	2
ASI-USC-TA1	2

Table 7: Number of trials annotated for each advisor condition.

experimental confederates, and (ii) we believed that there would be some level of phonetic entrainment between the participants in the 'human-advisor' condition and their human advisor, which would introduce an additional confounding variable into our analysis of phonetic entrainment. For the trials involving AI advisors, we sampled trials relatively equally across all six AI advisors. We sampled at the team level, so sampling an additional team for a given AI advisor results in two additional trials for that AI advisor (since each team completes two Minecraft missions).

We exclude trials that were for the purpose of training participants on how to perform the task. We disfavor—but do not completely exclude—trials with data quality issues (e.g. trials that are missing utterances due to technical issues with the audio capture setup). For trials in which the audio capture for one or more speakers failed due to technical issues, we were still able to annotate dialog acts, sentiment and emotion, but were unable to annotate for CLC events and entrainment.

B.2 Speaker Demographic

Speaker demographics are provided in Table 8.

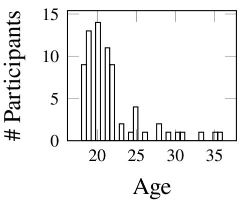
B.3 Annotator Demographic

Annotator demographics are provided in Table 9.

B.4 Annotator expertise

Our annotators have the necessary expertise to perform the annotation tasks. Four out of the five annotators are doctoral students that are 2–5 years into their PhD, working in areas that provide them a far greater level of expertise than can be found among crowdsourced annotators. Details on annotator expertise and training are provided in Table 10.

Annotators 1 and 2 are trained on the MRDA manual, a complex 129-page technical document (i.e., difficult to train crowdsourced annotators on).



(a) Age distribution

Sex	Count
Male	56
Female	12
Nonbinary	1
Prefer not to respond	2

(b) Sex

Ethnicity	Count
White/Caucasian	41
Asian	13
Hispanic or Latino	8
Middle Eastern	1
White/Caucasian, Hispanic or Latino	5
Hispanic or Latino, Other	2
White/Caucasian, Asian	2
White/Caucasian, Asian, Pacific Islander	1
White/Caucasian, Hispanic or Latino, Asian	1

(c) Ethnicity

English proficiency level	Count
Native or Bilingual Proficiency	59
Full Professional Proficiency	9
Professional Working Proficiency	4

(d) English proficiency

Highest education level achieved	Count
Some college/currently enrolled	48
12th grade	7
Some training in a master's program and/or graduated from a master's program	6
Graduated from college	5
Some training in a doctoral program and/or graduated from a doctoral program	1

(e) Education

Table 8: Aggregated speaker demographic data for selected dimensions.

Specification	Value
Age	23–33 years
Gender	Female (3), Male (2)
Race/ethnicity	East Asian (2), South Asian (2), Middle Eastern (1)
Native language	Korean (1), Tamil/Hindi/English (1), English (1), Persian (1), Sindhu/Urdu (1)
Socioeconomic status	Middle class (4), upper middle class (1)

Table 9: Annotator demograpics

Annotator #	Training	Annotation types
1	Undergraduate English major, took linguistics course, trained on MRDA manual.	Transcript correction, DA
2	PhD student in Computer Science working on NLP research, trained on MRDA manual	Transcript correction, DA
3	PhD student in Linguistics	Sentiment, Emotion, CLC
4	PhD student in Linguistics	Sentiment, Emotion, CLC, Entrainment
5	PhD student in Linguistics	Entrainment

Table 10: Annotator training

Annotators 3 and 4 are trained on CLC annotation, which involves a high level of inference and cognitive/working memory load. Additionally, the CLC annotation guidelines were developed by two other doctoral students and an NLP faculty member that performed an extensive review of existing CLC definitions and consulted with three external domain experts on CLC when developing the guidelines (the domain experts are mentioned in the Acknowledgments section which will be visible in the camera-ready version).

Annotators 4 and 5 used Praat to perform the Entrainment annotations. Praat is a specialized tool for speech analysis, and using it correctly requires expertise.

B.5 Speech Situation, Recording Quality

The audio recordings were conducted as part of a remote experiment that took place in 2022. Spoken, synchronous participant dialog was captured using the participants' own computers, often with background noises (which we try to annotate). The dialog was spontaneous, arising in the context of the collaborative virtual search-and-rescue task being performed by the participants. The intended audience for the speakers are their teammates that are performing the search-and-rescue task with them at the moment.

B.6 Database contents

The entirety of the MultiCAT dataset is provided through a single SQLite3 database (`multicat.db` in the supplementary material for the paper). The entity-relation diagram showing the structure of the database (tables, foreign key relationships, etc.) is shown in Figure 2.

C Items per class in MultiCAT

Tables 11, 12, 13, 14, 15, and 16 show the number of items per class in each task within MultiCAT. Note that some tasks allow multiple labels for a single utterance, so the number of items for a particular class in a task do not add up to the number of utterances annotated for that task.

D Breakdown of annotations by team and trial

The breakdown of annotations in MultiCAT by team and trial are shown in Table 17. Different tasks had different goals and different levels of complexity, so trials that were ideal for some were not always ideal

Class	Count
2	19
%	92
%-	123
%–	125
aa	1858
aap	10
am	14
ar	58
arp	1
b	39
ba	227
bc	6
bd	17
br	46
bs	17
bsc	94
bu	113
cc	1201
co	889
cs	251
d	206
df	233
e	449
fa	121
fe	152
ft	140
fw	1
g	58
j	44
m	136
na	263
nd	45
ng	32
no	43
qo	9
qr	52
qw	308
qy	808
r	44
s	6033
t1	141
x	116
z	264

Table 11: Items per class for DA classification

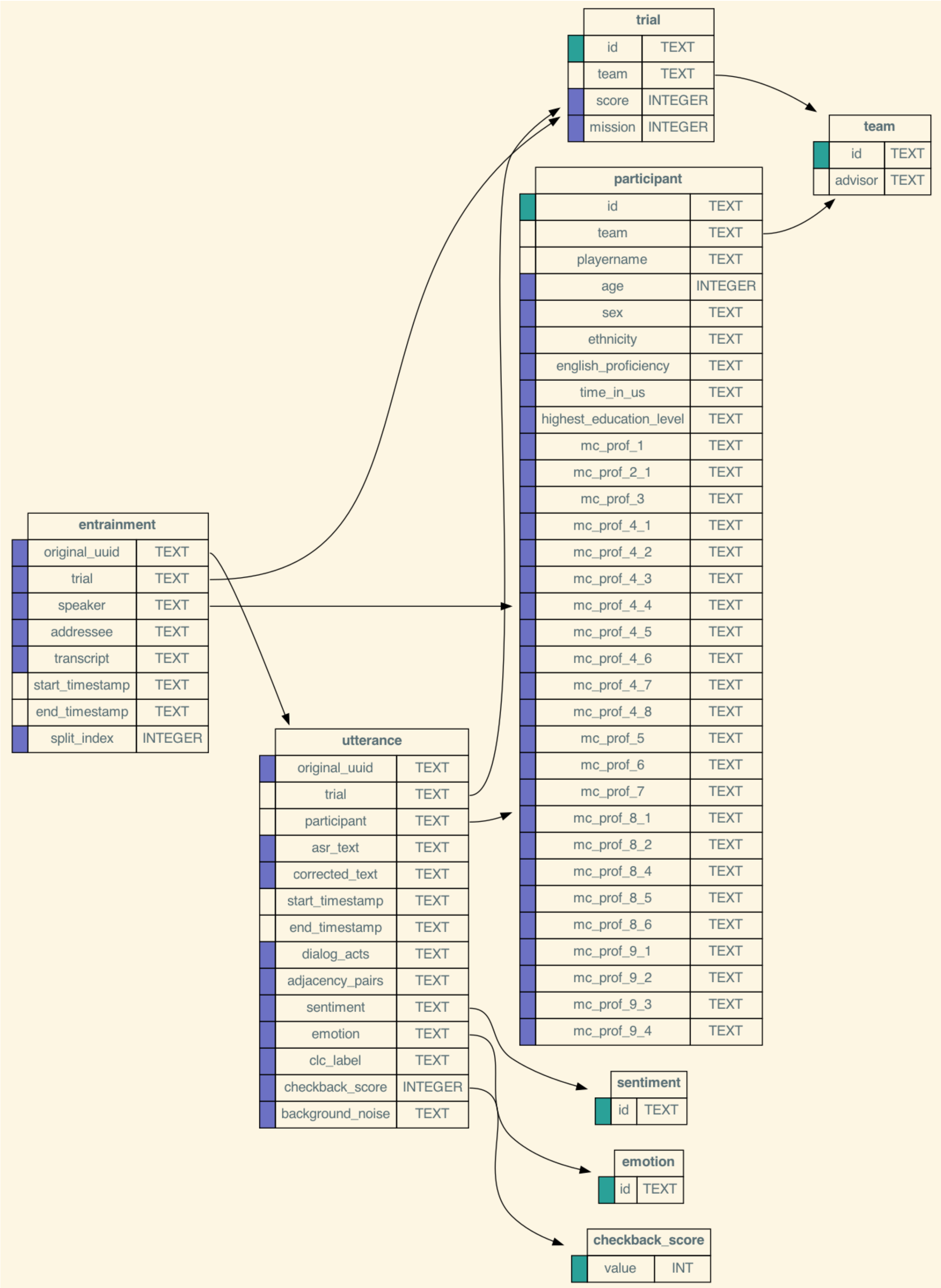


Figure 2: Entity-relation diagram for the MultiCAT database.

Class	Count
Neutral	4081
Positive	2436
Negative	1214

Table 12: Items per class for sentiment analysis

Class	Count
Neutral	5977
Joy	571
Sadness	452
Fear	319
Surprise	280
Anger	66
Disgust	66

Table 13: Items per class for emotion prediction.

Class	Count
a	4115
b	4473

Table 14: Items per class for adjacency pair identification.

Class	Count
a	3671
b	2767
c	386

Table 15: Items per class for CLC detection.

Class	Count
Addressee	2896

Table 16: Items per class for entrainment detection.

for all annotation types. For entrainment detection annotation, teams with two missions composed of clear audio files were selected. For sentiment and emotion annotation, extra trials were selected with the goal of increasing examples of small minority classes.

E Dataset comparison

Table 18 shows a comparison of MultiCAT to a number of relevant datasets.

F Feature engineering for the score prediction model

The features used for the score prediction results in § 7 are listed in Table 19.

G Model training details

Below are the details of parameters, computational resources used and specifics of our training procedures for our baseline models.

G.1 LLaMA Baseline

We provide LLaMA baseline results for DA, Sentiment, and Emotion classification tasks. For all the experiments, we use the instruction tuned 8B version of the model. To predict the label for an utterance, we provide 5 previous and 5 next utterances to serve as context. We fine-tune the models on the training set and report the results on the testset. Fine-tuning the model takes about an hour on one A100 GPU.

G.2 DA classification

The training, validation, and test splits we used are shown in Table 20. We use version 1.13.1+cu117 of the PyTorch library (Paszke et al., 2019). The learning rate is set to 10^{-4}. The AdamW optimizer (Loshchilov and Hutter, 2019) is used with a decay of 10^{-5}. We train for a maximum of 100 epochs with early stopping after no improvement on the validation set for 10 epochs. The model has around 127M parameters, and takes $\approx$ 23 minutes to train. All experiments are performed on a single NVIDIA RTX A6000 GPU.

G.3 CLC detection

For the logistic regression model, we use as the training set the following 25 trials: T000603, T000604, T000607, T000608, T000613, T000627, T000628, T000631, T000632, T000633, T000634, T000635, T000636, T000637, T000638, T000713, T000714,

Team	Trial	SentEmo	CLC	DA	AP	Entrainment
TM000201	T000602	✓				
TM000202	T000603	✓	✓	✓	✓	✓
TM000202	T000604	✓	✓	✓	✓	✓
TM000203	T000605	✓	✓	✓	✓	✓
TM000203	T000606	✓	✓	✓	✓	✓
TM000204	T000607	✓	✓	✓	✓	
TM000204	T000608	✓	✓	✓	✓	
TM000205	T000609	✓		✓	✓	
TM000205	T000610	✓		✓	✓	
TM000206	T000611	✓		✓	✓	
TM000206	T000612	✓		✓	✓	
TM000207	T000613	✓	✓	✓	✓	
TM000207	T000614	✓				
TM000210	T000619	✓				
TM000210	T000620	✓		✓	✓	
TM000211	T000621	✓				
TM000211	T000622	✓		✓	✓	
TM000212	T000623	✓	✓	✓	✓	
TM000212	T000624	✓		✓	✓	
TM000213	T000625	✓		✓	✓	
TM000213	T000626	✓		✓	✓	
TM000214	T000627		✓	✓	✓	
TM000214	T000628	✓	✓	✓	✓	
TM000216	T000631	✓	✓	✓	✓	
TM000216	T000632	✓	✓	✓	✓	
TM000217	T000633	✓	✓	✓	✓	
TM000217	T000634	✓	✓	✓	✓	
TM000218	T000635	✓	✓	✓	✓	
TM000218	T000636	✓	✓	✓	✓	
TM000219	T000637	✓	✓	✓	✓	
TM000219	T000638	✓	✓	✓	✓	
TM000236	T000671	✓	✓	✓	✓	
TM000236	T000672	✓		✓	✓	
TM000252	T000703		✓	✓	✓	
TM000252	T000704		✓	✓		
TM000257	T000713	✓	✓	✓	✓	
TM000257	T000714	✓	✓	✓	✓	
TM000258	T000715	✓	✓	✓	✓	
TM000258	T000716	✓	✓	✓	✓	
TM000260	T000719	✓	✓	✓	✓	✓
TM000260	T000720	✓	✓	✓	✓	✓
TM000262	T000723	✓	✓	✓	✓	✓
TM000262	T000724	✓	✓	✓	✓	✓
TM000264	T000727	✓	✓	✓	✓	
TM000264	T000728	✓	✓	✓	✓	
TM000265	T000729	✓	✓	✓	✓	
TM000265	T000730	✓	✓	✓	✓	
TM000269	T000737	✓	✓	✓	✓	
TM000269	T000738	✓	✓	✓	✓	

Table 17: A list of all trials with the team that trial represents indicating which types of annotation each trial contains.

Dataset	Natural	Multiparty	Audio	Task†	AP	DA	CLC	Sent.	Emo.	Entrainment
MRDA (Shriberg et al., 2004)	✓	✓	✓		✓	✓				
SwDA (Jurafsky et al., 1997)	✓		✓			✓				
STAC (Asher et al., 2016)	✓	✓		✓	✓*					
TRAINS (Heeman and Allen, 1995)	✓		✓	✓						
DBOX (Petukhova et al., 2014)			✓	✓		✓				
DailyDialog (Li et al., 2017)						✓			✓	
Ubuntu (Lowe et al., 2015)	✓			✓						
DeliData (Karadzhov et al., 2023)	✓			✓						
SIMMC (Kottur et al., 2021)				✓		✓				
RAVDESS (Livingstone and Russo, 2018)			✓							
GEMEP (Bänzinger et al., 2012)			✓						✓	
IEMOCAP (Busso et al., 2008)			✓						✓	
YouTube (Morency et al., 2011)	✓		✓					✓		
ICT-MMMO (Wöllmer et al., 2013)	✓		✓					✓		
CMU-MOSEI (Bagher Zadeh et al., 2018)	✓		✓					✓	✓	
MELD (Poria et al., 2019)		✓	✓					✓	✓	
Columbia Games Corpus	✓		✓	✓	✓*	✓				✓
Brooklyn Multi-Interaction Corpus (Weise et al., 2022)	✓		✓						✓	✓
Suicide Risk Assessment Corpus (Baucom et al., 2014)	✓		✓	✓						✓
Couples Therapy Corpus (Christensen et al., 2004)	✓		✓	✓						✓
Fisher Corpus (Cieri et al., 2004)			✓							✓
MultiCAT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

† Task = Task oriented
* STAC and the Columbia Games Corpus include discourse structure, but not using APs.

Table 18: Comparison of MultiCAT with related datasets.

Feature set	Feature name	Feature description
Proficiency	avg_mc_prof_2_1	Self-reported confidence for learning and succeeding at a new video game or set of game-related skills after minimal practice
	avg_mc_prof_4_1	Self-reported confidence learning the layout of a new virtual environment
	avg_mc_prof_4_2	Self-reported confidence for communicating their current location in a virtual environment to members of a team
	avg_mc_prof_4_3	Self-reported confidence for coordinating with teammates to optimize tasks
	avg_mc_prof_4_4	Self-reported confidence for maintaining an awareness of game/task parameters (e.g., time limits, goals, etc)
	avg_mc_prof_4_5	Self-reported confidence for Learning the purposes of novel items, tools, or objects
	avg_mc_prof_4_6	Self-reported confidence for remembering which places they have visited in a virtual environment
	avg_mc_prof_4_7	Self-reported confidence for controlling the movement of an avatar using the W, A, S, and D keys + mouse control
	avg_mc_prof_4_8	Self-reported confidence for keeping track of where they are in a virtual environment
	avg_mc_prof_9_1	Number of years using a computer for any purpose
	avg_mc_prof_9_2	Number of years using a computer to play video games
	avg_mc_prof_9_3	Number of years using a system other than a computer to play video games (e.g., mobile phone, gaming console, arcade console)
	avg_mc_prof_9_4	Number of years playing Minecraft (any versions or styles of play)
Emotion	emo_neutral	Number of utterances in the trial labeled with the 'neutral' emotion.
	joy	Number of utterances in the trial labeled with the 'joy' emotion.
	surprise	Number of utterances in the trial labeled with the 'surprise' emotion.
	sadness	Number of utterances in the trial labeled with the 'sadness' emotion.
	disgust	Number of utterances in the trial labeled with the 'disgust' emotion.
	anger	Number of utterances in the trial labeled with the 'anger' emotion.
	fear	Number of utterances in the trial labeled with the 'fear' emotion.
Sentiment	sent_neutral	Number of utterances in the trial labeled with the 'neutral' sentiment.
	positive	Number of utterances in the trial labeled with the 'positive' sentiment.
	negative	Number of utterances in the trial labeled with the 'negative' sentiment.
AP	neither	Number of utterances in the trial that have neither a or b AP annotations.
	b	Number of utterances in the trial that only have b annotations.
	a	Number of utterances in the trial that only have a AP annotations.
	both	Number of utterances in the trial that have both a and b AP annotations
CLC	clc_none	Number of utterances in the trial that do not have CLC labels.
	clc_a	Number of utterances in the trial that only have a CLC annotations.
	clc_b	Number of utterances in the trial that only have b CLC labels.
	clc_c	Number of utterances in the trial that only have c CLC labels.
	clc_ab	Number of utterances in the trial that have both a and b CLC labels.
	clc_ac	Number of utterances in the trial that have both a and c CLC labels.
	clc_bc	Number of utterances in the trial that have both b and c CLC labels.
DA	s	Number of utterances in the trial that only have 's' labels.
	qr	Number of utterances in the trial that only have 'qr' DA labels.
	qw	Number of utterances in the trial that only have 'qw' DA labels.
	qy	Number of utterances in the trial that only have 'qy' labels.
	x	Number of utterances in the trial that only have 'x' labels.
	z	Number of utterances in the trial that have 'z' labels.
	qy_s	Number of utterances in the trial that have both 'qy' and 's' labels.
	qw_s	Number of utterances in the trial that have both 'qw' and 's' labels.
	qw_qy	Number of utterances in the trial that have both 'qw' and 'qy' labels.
	qw_qy_s	Number of utterances in the trial that have 'qw', 'qy', and 's' DA labels.
	qr_s	Number of utterances in the trial that have both 'qr' and 's' labels.
	qr_qy_s	Number of utterances in the trial that have 'qr', 'qy', and 's' labels.
	qo_s	Number of utterances in the trial that have both 'qo' and 's' labels.

Table 19: 'qr_qw_s' and 'qo' are omitted since they are not there in mission one. For the items in the 'Proficiency' feature set, the values are averages across all the teammates in a particular trial. All self-reported confidence values are on a scale of 0–100.

Split	# of trials	Trial IDs
Train	28	T000603, T000604, T000611, T000612, T000620, T000622, T000623, T000624, T000627, T000628, T000631, T000632, T000635, T000636, T000637, T000638, T000703, T000704, T000713, T000714, T000715, T000716, T000719, T000720, T000723, T000724, T000729, T000730
Validation	5	T000613, T000607, T000608, T000633, T000634
Test	12	T000605, T000606, T000671, T000672, T000625, T000626, T000727, T000728, T000737, T000738, T000609, T000610

Table 20: Train, validation, and test split composition for the DA classification and AP detection tasks.

T000715, T000716, T000719, T000720, T000723, T000724, T000729, T000730.

For the check-back detection step, we used the following 20 trials as the training set: T000603, T000604, T000627, T000628, T000631, T000632, T000635, T000636, T000637, T000638, T000713, T000714, T000715, T000716, T000719, T000720, T000723, T000724, T000729, T000730, and the following 5 trials as the validation set: T000607, T000608, T000613, T000633, T000634.

The detection of the call-out step with the logistic regression model takes 0.1 second to train.

We adopted the Transformer-based RoBERTa-base model for the detection of the check-back step. The learning rate is set to 5×10^{-5}, the model is trained with a batch size of 16 for 3 epochs. This model takes approximately 30 minutes to train.

The CLC detection experiments are performed on a Apple M1 CPU.

G.4 Sentiment and emotion classification

We train our sentiment and emotion baseline on a high performance computing environment with a Tesla V100S-PCIE-32GB GPU. For the sentiment and emotion classification tasks, we use the same training, validation, and test splits as in Table 20, except for including an additional trial (T000614) in the validation split.

We train this model using version 2.2.0+cu121 of PyTorch. Our baseline model contains 1,904,690 parameters. Our best hyperparameter settings are a learning rate of 10^{-3} with an Adam optimizer with a weight decay of 10^{-4}.

We perform a limited grid search over our pre-training corpora, then fine-tune with MultiCAT data on the best of these. The model takes approximately 15 minutes to train and 2 minutes for fine-tuning.

G.5 Entrainment identification

We train our entrainment model using PyTorch version 1.9.0+cu111 with torchaudio version 0.7.0[6] and NumPy version 1.22.4 (Harris et al., 2020). This is done on an NVIDIA A100-PCIE-40GB GPU. We use the same hyperparameters identified as best in Nasir et al. (2020). Training of the model on the Fisher corpus took an average of 70 minutes, and testing on MultiCAT takes 3.5 minutes (for all 30 iterations).

G.6 Score prediction

We evaluate ridge regression models for score prediction in § 7. We use the implementation of ridge regression in scikit-learn v1.4.0 (Pedregosa et al., 2011), with the L_2 regularization coefficient $\alpha = 10$. This hyperparameter was selected using a manual coarse-grained grid search between 0.1 and 50, such that the value of the mean MAE across folds and the standard error of the mean were minimized for the Mission 2 results. Experiments were carried out on a 2021 MacBook Pro with an Apple M1 Max CPU. The results in Table 6 were generated by a script that took approximately 4 seconds to execute—including both data loading and model training.

H Software

The code used to generate the database and the results in the paper will be added to the supplementary material for the camera-ready version upon paper acceptance.

I ASR transcript correction guidelines

Basic Setup The data should be in CSV format with one column for ASR and one column of corrected transcripts. The annotator is expected to listen to the full audio and read the ASR transcripts, whenever there are any discrepancies, those should be corrected and entered only in the corrected transcripts column.

Segmentation The segmentation of speaker utterances as done by ASR is not to be changed. For example, even if the annotator feels utterance B

[6] https://pytorch.org/audio/stable/index.html

should come before utterance A, they should not change the order of the utterances.

Missing Utterances At times the ASR fails to pick up on small utterances, especially those that are just a few words long. In that case, a new row should be inserted in the CSV file and the text of the utterance should be manually entered. The field for the ASR transcript should be left empty. The annotator should also enter the speaker name and start and end timestamps.

Relative Order of New Utterance The utterance should be inserted based on the start timestamp and its relative order with the already present utterances.

Noise Picked up by ASR When ASR picks up noise as an utterance, a special character of hyphen "-" should be added as the corrected transcript.

J DA annotation guidelines

J.1 MRDA Framework

Our annotations follow the same guidelines as that of the ICSI MRDA corpus. The manual for MRDA contains detailed examples and definitions of different tags. This manual further builds on the MRDA manual (Dhillon et al., 2004) and addresses special cases we encountered when annotating MultiCAT.

J.2 Questions

Discontinuous Question When speaker A asks a question but they get interrupted by speaker B. after the interruption, speaker A goes on to finish the question. Two scenarios can arise.

- Speaker B answered the question, in this case the subsequent utterances by speaker A would be marked with statement general tag and elaboration specific tag. Since speaker A's intent behind the latter utterances is not to elicit an answer. Check page 34 of MRDA manual for a similar use case.

- Speaker B does not answer the question, the rest of speaker A utterances completing the question would get the same question tag(s).

J.3 Segmentation with Pipe

Floor Mechanisms (FM) <fg>, <fh>, <h> at the start or end of an utterance can be ignored. No need to pipe separate an utterance or include the FM tag in the label.

Short Response For tags <aa> and <ar> at the start or end of an utterance, make the response tag as part of a single combined utterance tag. That is, the general tag will be shared by the whole utterance.

Different General Tags with Pipe Pipe should be used for cases where segments of the utterance require different tags and cannot be merged into one label because of different general tags. The pipe would then be added to both the utterance and the label.

Utterance	DA
Oh you do? |So you probably discard	qh | s^cs

Table 21: An example illustrating the use of pipe bar to annotate an utterance for multiple general tags.

J.4 Acknowledgment <bk> & Accept <aa>

<bk> and <aa> tags have been merged into a single tag - <aa>.

J.5 <df> and <e> for a Single Utterance

The tag <df> can be assigned to a single utterance without having to associate it with a previous utterance. The same is not true for <e>. <e> tag can only be assigned in relation to some previous utterance.

Special case of <df> and <e> in same utterance If an utterance were to be segmented to assign <df> tag while some portion has already been assigned the <e> tag, the <df> and <e> tags can be merged under the same general tag (if after pipe <df> was to receive the same general tag as well)

Speaker	Utterance	DA
A	So yeah I would move.	s^cs
B	Um.	h
A	down to Breaker's Bridge and shore it up, cause I don't think there's anything we can do.	s^df^e

Table 22: <df> and <e> can occur in the same utterance but <e> still has to be in relation to a prior utterance of the same speaker.

J.6 Commitment <cc> in Present Actions

In MultiCAT data, players often verbalize the action they are carrying out at the present moment, any such actions should also be considered as <cc>.

Utterance	DA
yep on my way.	s^aa^cc

Table 23: <cc> for present actions.

K Sentiment/emotion annotation guidelines

One task to complete during this summer's annotation effort is the annotation of utterances for sentiment and emotion. This document discusses the method that should be used when annotating each.

K.1 Key terminology

K.1.1 Utterance

For purposes of this task, we define the term **utterance** as a single unit transcribed by Google's ASR. In some cases, this will correspond to a single sentence without a pause; in others, this may actually be composed of more than one sentence. Occasionally, a single sentence is even split into two utterances by the ASR.

K.1.2 Emotion

Emotion in this task refers to the discrete emotion shown by a speaker during an utterance. The emotion is selected from the set of labels described in section 3 below.

K.1.3 Sentiment

Sentiment in this task refers to the feelings a speaker shows towards the topic of an utterance. The sentiment may be positive, negative, or neutral. Sentiment labeling is discussed in section 4 below.

K.2 Basic annotation procedure

You will be asked to make your annotations using spreadsheets and while accessing the full audio files for a mission. Below is the annotation procedure that we will be following.

K.2.1 Materials needed

To complete this annotation task, you will need a spreadsheet containing each of the corrected/uncorrected utterances (which should be provided to you) with empty columns where you will enter your annotation labels, as well as the corresponding audio files.

You should select a quiet place to work and use headphones to ensure that you can clearly hear the entirety of the audio.

K.2.2 Procedure

For this task, you should have the transcript and label spreadsheet open while listening to the audio. If you cannot look at the transcript and listen to the audio at the same time, you should read the transcript for each single utterance immediately before listening to that utterance.

For the sake of consistency, we will be using **uncorrected** transcripts for this task. This means that the words may not form a logical sentence, and at times may be difficult to understand. When this happens, do your best to pay attention to the words in the recording (as these should make sense) and use these to help inform your decisions.

You will need to download the transcripts and the relevant audio files from Hidden for double-blind peer review. The transcripts may be found in the following location: Hidden for double-blind peer review. The audio files may be found in: Hidden for double-blind peer review. Some of these transcript files may contain corrected transcripts; however, you should focus on the uncorrected transcripts (the column labeled 'utt' or 'utterance').

Select a transcription and the corresponding audio; open the transcription to take up at least half of your screen, ensuring that you can see the entirety of each transcribed utterance that is within the window.

After listening to a single utterance, pause the recording, then enter the emotion label and the sentiment label into the corresponding cells in the spreadsheet. You may then play the recording again and examine the next utterance.

K.3 Emotion task

The first of the two annotations that you will be completing as you go through the files is the emotion task. For this task, you will need to decide which of a set of emotions is the best label for each individual utterance, as defined above. The set of labels used in this task and examples of annotations for each appear below.

K.3.1 Emotion labels

While there are several methods for capturing emotional information from audio, we are using a set composed of Ekman's universal emotions + a neutral label. This label set is:

1. **anger**: the speaker is angry, upset, and reveals this through words, tone or both.

2. **disgust**: the speaker is disgusted; in this dataset, disgust frequently appears when a player walks into the same trap room more than once, when someone is having a little bit of trouble with the controls, or when any sort of glitch occurs. This emotion label is more like frustration than anger.

3. **fear**: the speaker is afraid of something.

4. **joy**: the speaker is happy, having a good time, or otherwise enjoying something. This emotion frequently occurs at the end of missions immediately after time has run out, though some speakers show moments of joy throughout the mission.

5. **neutral**: (no clear emotion)–the speaker doesn't demonstrate any emotions; they may be explaining something or providing information about their movements to their team. This sort of neutral language is very common in the ASIST data.

6. **sadness**: the speaker is sad or disappointed, often because something has happened that they did not want to have happen (like repeatedly entering a trap room), or because something hasn't happened that they wanted to see happen (e.g. the number of victims saved is lower than they had hoped).

7. **surprise**: something surprising has happened, the speaker is suddenly given new unexpected information or corrected about something they thought they knew but that turned out to be incorrect.

Each utterance should be given a single label. This label may be based on the words that the participant produces, the way in which they speak, or both.

K.3.2 How to decide which emotion label to select

Determining which label to use is often straightforward; sometimes, however, you may not be sure of which label to assign an utterance. In general, follow these rules:

1. If an utterance contains no obvious emotional information, give it a label of neutral

2. If most of an utterance contains no obvious emotional information, but one part of it does contain emotion, provide the label of the non-neutral emotion demonstrated

3. If an utterance contains two emotions, do the following:

   - If one emotion seems much stronger than the other, choose the stronger emotion
   - If one emotion dominates the utterance, choose the dominant emotion
   - otherwise (assuming equal parts of each of two emotions):
     (a) If one emotion is fear and the other is anything else, choose fear
     (b) If one emotion is sadness and the other is anything but fear, choose sadness
     (c) If one emotion is anger and the other is not fear or sadness, choose anger
     (d) If one emotion is disgust and the other is joy or surprise, choose disgust
     (e) If one emotion is joy and the other is surprise, choose joy
   - If there are ever three emotions in one utterance, follow the points above to make your decision about which to select

K.3.3 Examples of emotion annotations

"Okay can you make sure you mark it?" Said with a neutral tone, this would be given the label neutral. The speaker is making a request of another player.

"Oh shoot that's the wrong one" The participant suddenly realized they have gone to the wrong location. This should be given the label surprise.

"and then wacky fun little update guys both of our C zones are blocked right now" While the ASR transcription isn't perfectly accurate, this speaker is indicating that they are stuck in a room. With the intonation from the audio, we can tell that 'wacky fun little update' is sarcastic, so this utterance should be given the label disgust.

"shit" This speaker just shouted this word out, showing that they were feeling mad, this would be given the label anger.

"guys I'm starting to think we're not going to get everyone" This speaker is disappointed that their performance is not as good as the team had hoped. This would be given the label sadness.

"I was like 3 seconds away oh I died" At the end of the game, the speaker has not managed to save the last victim they were carrying. Then the game ends by showing the speaker's character dying. Without the audio, it may seem as though this person is disgusted, angry, or surprised, but they are in fact laughing and having fun, while being surprised by the event. This could have been labeled either joy or surprise, so following the guidelines above, we select label joy.

"Ah, what's happening?" The mission has ended and the screen has suddenly changed, but the speaker thinks they have done something wrong somehow. They show both surprise and fear, so using the guidelines above, we select the label fear.

"oh geez now she's been a red turn its meeting throws a 720" While the ASR is not quite right, this person is annoyed at an aspect of the mission that they have no control over (their speed). This could show surprise, disgust, or anger, so using the guidelines above we select anger.

K.4 Sentiment

The second annotation task that you will complete while going through these files is sentiment annotation. For this task, you will assign each item a sentiment label according to the sentiment expressed in the statement. For this task, as with the above, you will want to pay attention to both what is said and how it is said.

K.4.1 Sentiment labels

Sentiment: the content/meaning of each utterance should be marked as one of the following.

1. **positive**: the utterance refers to a subject that the speaker feels positively about.

2. **neutral**: the utterance does not reveal positive or negative sentiment; this is generally the case with instructions, updates, descriptions of players' movements and when speakers provide general information.

3. **negative**: the utterance refers to a subject that the speaker feels negatively about.

K.4.2 How to decide which sentiment label to select

Because there are only three sentiment labels to select from, it is much less likely that you will have to make difficult decisions about which to choose.

1. If there is no indication of either positive or negative sentiment, choose the neutral label

2. If any part of the utterance demonstrates positive or negative sentiment, select that sentiment, even if the majority of the utterance is neutral

3. If both positive and negative sentiment are shown in equal amounts in the same utterance, select the negative label

4. Politeness does not convey any information other than politeness. Thus, select neutral label

5. 'Okay' should be labeled depending on tone and pitch

    - negative: sarcasm, annoying situation
    - neutral: gap filler
    - positive: other than the aforementioned

There is a correlation between sentiment labels and emotion labels (e.g. 'happy' utterances would tend to also have a positive sentiment), although there is not an exact mapping of sentiments onto emotions (e.g. 'surprise' could be positive or negative). The vast majority of the utterances seem to be neutral in both emotion and sentiment, and that's okay. One of the recordings I listened to only had one utterance that showed a non-neutral emotion/sentiment value (the last utterance, actually).

Sometimes, however, the emotion a participant shows is NOT the same as the sentiment they express. For example, sometimes someone expresses joy through their tone, but the words they are saying actually indicate a negative sentiment (e.g. they are having fun playing the game, but they say 'We did really poorly this round!').

K.4.3 Examples of sentiment annotations

"It might actually be best to start in the middle and then work our way either left or right because the middle is where we spawn" This speaker is giving suggestions on what they think is the correct way to organize their movements during a mission that is just starting. They are neutral in their tone. This should be labeled neutral.

"Okay engineer to enter so critical in here yeah" The ASR has not given an accurate transcription here, but we can see that most of the words themselves seem neutral. However, with the speaker's tone, we see that they feel positively about the event taking place at the end (where a critical victim is found), so this would be labeled positive.

"Other that sorry that's the one you know it's not okay so we got that b there's two critical Zone here speak out that one but" The ASR is again not quite accurate, but we can see that this person does not seem to feel positively about the room that they have just entered. Using this knowledge, plus phrases like 'sorry' and 'it's not okay', this would be labeled as negative.

L Entrainment annotation guidelines

In this annotation task, we search for the intended listener of a given spoken unit. You task is to listen to the audio, read the transcripts for every utterance in the recording, find the inter-pausal units within each utterance, and ascertain who the inter-pausal unit is aimed at.

L.1 Key terminology

L.1.1 Utterance

A section of the spoken interaction that the automatic transcription service has detected as a unit of speech.

L.1.2 Vocal Entrainment

Vocal Entrainment is the shift in vocalic features (such as fundamental frequency) of a speaker in order to resemble their conversation partner.

L.1.3 Inter-pausal Unit (IPU)

A stream of audio separated by a pause of 50ms or more. This can be a whole or part of an utterance.

L.1.4 Split indices

Entrainment task works at the IPU level. Many utterances in this dataset will have pauses longer than 40ms within them (i.e they contain multiple IPUs that have the same UUID). They will need to be split up. The resultant chunks will be assigned split indices (0,1,2...) and will retain their parent utterance's UUID. These split indices ensure that all splits of a given utterance retain their original metadata.

L.2 Basic annotation procedure

For this task, you will be working to assess and correct the IPU boundaries on a automatically filled Praat textgrid. For each IPU you correct and finalize, you will add the corresponding transcription in the 'silences' tier from the transcript spreadsheet provided. Finally, you will identify the intended addressee of every IPUs and annotate for it in the 'addressee' tier. Your final submission is a corrected textgrid with labels in the 'silences' and the 'addressee' tiers.

You will be asked to make your annotations using spreadsheets and the audio files from the individual recording channels for each player in given a mission. The procedure is outlines in the 'Procedure' section below.

L.2.1 Materials and technology needed

- Praat software.
- The spreadsheet containing the corrected utterances for a given trial.
- The corresponding audio files.
- Automatically filled textgrids (one per audio file) with two tiers, 'silences' and 'addressee'. The 'silences' tier will have two types of automatically detected labels: 'silence' (which is the label for non-speech sounds as well as silences), and 'sound' (for speech).

  You should select a quiet place to work and use headphones to ensure that you can clearly hear the entirety of the audio.

L.2.2 Procedure

For this task, keep the transcript open on any spreadsheet reader, along with the audio and Praat textgrid open on Praat.

1. Download the transcripts, textgrids and the relevant audio files from Hidden for double-blind peer review. The transcripts may be found in the following location: Hidden for double-blind peer review, and the audio and textgrids in Hidden for double-blind peer review.
2. On Praat, move your cursor to the first chunk where the experiment participant is speaking.
3. Listen until you hear the speaker pausing, and check if the pause is over 50 ms. You can see the length of the selected audio above the waveform, or by clicking on 'Query' > 'Get length of selection' in the menu on the top left corner of the screen. If the pause is less than 50 ms, continue listening until you hear a pause.
4. If you see a longer pause, make sure the start and end of the speech has boundaries on both the 'silences' and 'addressee' tiers. Drag the boundaries until they enclose the speech and

move them as close to the speech chunk as possible.

5. Ensure that the silences on each side of the speech chunk have the automatically generated label ‘silence’.

6. From the spreadsheet, copy and paste the chunk of the transcript that matches the words you hear into the ‘silences’ tier. These words may be just a portion of the utterance in the cell. The rest may belong to the following IPU.

7. Identify the addressee of the IPU. You can determine this from the context of the conversation. For example, the speaker could have called out to a specific player. Or the IPU could be part of an answer to a question asked in a previous utterance.

8. Add an addressee label in the ‘addressee’ tier. You have four options. If you identify a distinct addressee, annotate with the name of any one Minecraft roles played by the players (‘engineer’, ‘transporter’, ‘medic’).

9. Or, if you can’t identify a specific addressee, or if the IPU is directed at the experimenter, simply mark it as ‘all’.

10. Continue scrolling through the IPUs until you have corrected, transcribed and addressee-identified each IPU. Save your annotated textgrid frequently.

L.2.3 An example for IPU detection

Figure 3 has a Praat window open with the waveform (top), spectrogram (middle), as well as the textgrid (bottom) containing the automatically detected voice activity for the files ‘HSRData_ClientAudio_Trial-T000719_Team-TM000260_Member-E000888_CondBtwn-ASI-UAZ-TA1_CondWin-na_Vers-1.wav’ and ‘HSRData_ClientAudio_Trial-T000719_Team-TM000260_Member-E000888_CondBtwn-ASI-UAZ-TA1_CondWin-na_Vers-1.TextGrid’. The view shows the audio divided into chunks of sound and silence (labelled in the first tier). In reality, this is one inter-pausal unit in which the consonants have been incorrectly labelled as silences by the automatic speech detector. Our first task is to correct the IPU boundaries and add the transcript corresponding to it.

First, we remove the unwanted boundaries and labels such that only the initial and final boundaries remain. Next, we adjust the start and end boundaries until they enclose only speech. Finally, we add the text from the transcription spreadsheet. The end result should look like Figure 4.

L.2.4 An example for addressee identification

Using the same IPU as the above section, we now move on to identifying the speaker and their addressee. First, we look in the transcript spreadsheet for utterances preceding the IPU of interest, and who was the speaker. In the example, the utterances preceding ‘this is transporter there’s a critical victim in A4’ (‘this is’ and ‘three’) are also uttered by the same speaker (‘transporter’). By scrolling back (or zooming out, as seen in Figure 5 on the textgrid, you can see that both the previous utterances did not have a specific addressee (thus labelled ‘all’). Based on the context, we will mark this IPU as ‘all’ in the ‘addressee’ tier on the textgrid.

This completed the annotation task for this IPU, and we can scroll to the next one.

M CLC annotation guidelines

This document discusses the method of annotating closed-loop communication events in multi-party dialogues.

M.1 Definition of Closed-Loop Communication

In team communication, especially in emergency situations, there’s a standard scheme of communication, called Closed-loop communication. Closed-loop communication aims to achieve safe communication by reducing the risk of miscommunication and ensuring clear communication. Closed-loop communication is usually trained and adopted in high-stakes team environments like Crew Resource Management, medical surgery teams, and emergency departments. In our Minecraft games which simulate the urban search and rescue scenario, the appearance of Closed-loop communication is considered a good approach to team communication, although the participants of the game are not trained in doing so.

Closed-loop communication includes three phases:

Call-out The sender initiates a message.

Check-back The receiver acknowledges the message, usually by paraphrasing or repeating the

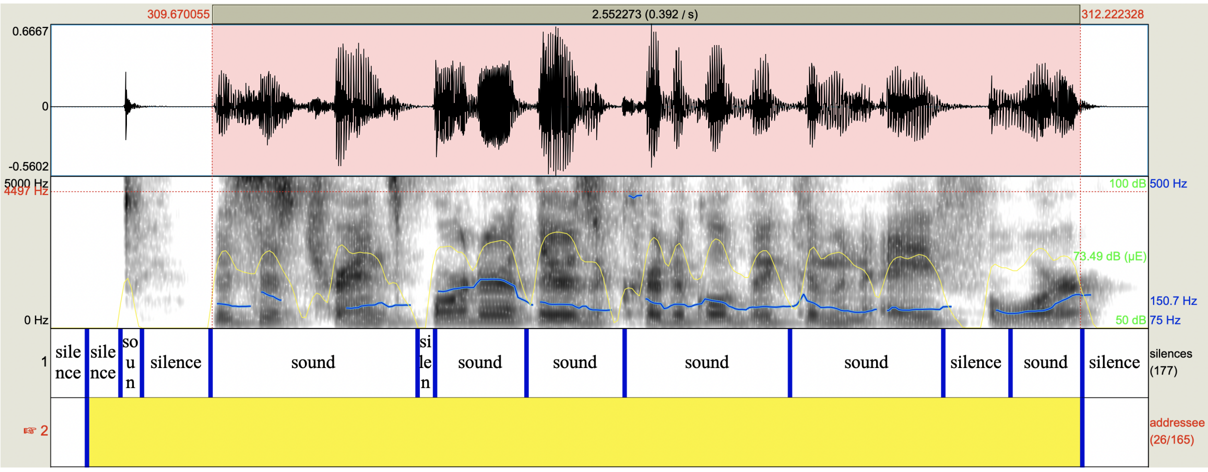


Figure 3: Original textgrid

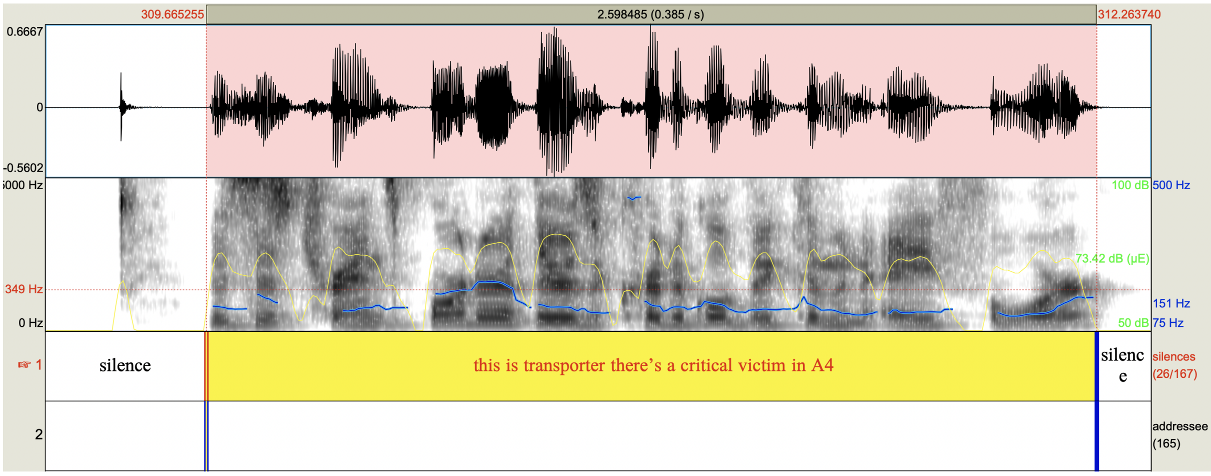


Figure 4: Textgrid with IPU boundaries and transcript

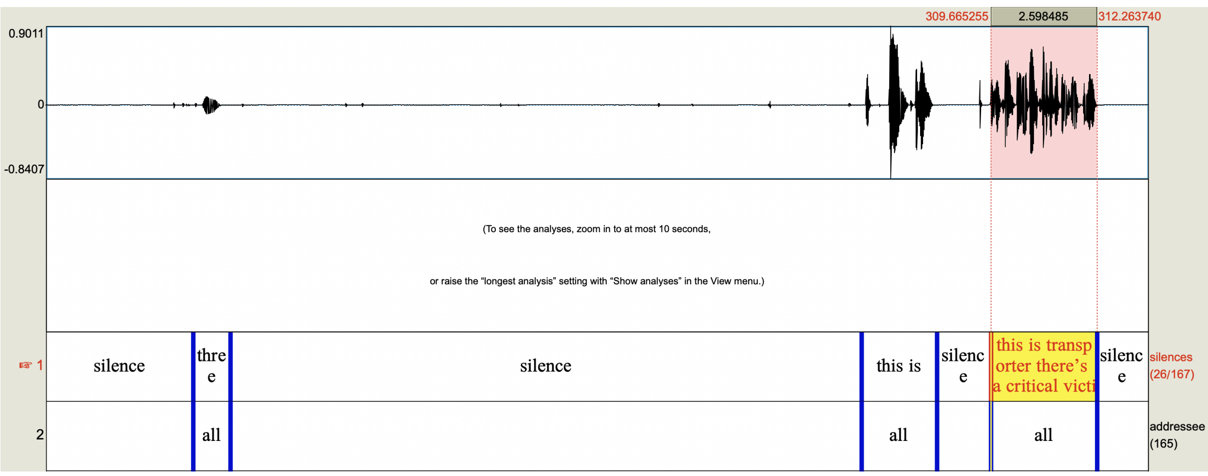


Figure 5: Textgrid with IPU boundaries and transcript

main information of the message.

Closing-of-the-loop The sender verifies that the message has been received and interpreted correctly.

Table 24 is an example of closed-loop communication. The detection of Closed-loop communication will be triggered by recognizing the Call-out phase, and then searching for the Check-back phase, and finally the closing-of-the-loop phase. There might be situations where only a sender calls out but no one checks back to the sender, or there're call-out and check-back but no final acknowledgment to close the loop. We have different labels for the three phases. Table 25 is a list of common semantic types of the CLC phases.

Role	Utterance	CLC Phase
Green	This is Green. I'm finishing this side, blue, could you check the central?	Call-out
Blue	This is Blue. I'll go check the central.	Check-back
Green	Thank you, Blue.	Closing-of-the-loop

Table 24: An example of the closed-loop communication

M.2 Labels and Scores

The transcripts of utterances are saved in CSV files. The annotations are in columns: CLC_Label, Checkback_Score.

At the beginning of each trial, there are several pre-game chatting utterances, which happen before players enter the scene and they were communicating with each other about team strategies. At the end of each trial, there're also several post-game utterances after the game session ends. We will not include those in our CLC annotation.

The three phases of the CLC are labeled with letters *a*, *b*, and *c*:

- Call-out: *a*
- Check-back: *b*
- Closing-of-the-loop: *c*

We follow the MRDA (Multi-Dimensional Annotation) framework for annotating adjacency pairs and adapt it to our CLC annotation with the format: **<CLC number><CLC phase>.<CLC number><CLC phase>-<nth speaker>[+...]**

The <CLC number> is the index number of CLC events, which helps us keep linking call-outs and their follow-up check-backs and closing-of-the-loops, especially when they are several utterances away from the call-outs. The <CLC phase> are *a*, *b*, and *c* phases for each CLC event. The <nth speaker> is useful when there're multiple check-outs for one call-out, and the [+...] suffix is used to note a continued CLC phase from the same speaker, which usually happens when a sentence is cut off into more than one utterances. For example:

8a.9a indicates two call-out events in one utterance, see table 27.

a+/a++ indicates continued call-out events by the same speaker, see table 28.

b+/b++ indicates the same person check-back to one call-out event, see table 29.

b-1/b-2 indicates two check-backs from different speakers to one call-out, see table 30.

The three phases are not necessarily closely next to each other. There might be some other utterances that insert between call-out and check-back, and check-back and closing-of-the-loop.

In our scripts, sometimes, the time span of each utterance might overlap, and starting timestamp may not be ordered properly. We need to pay special attention to the timestamps in order to make sense of the flow of conversations.

The **Checkback_Score** measures the quality of the check-back phases. If the check-back utterance repeated the key information in the call-out utterance, and shows the full understanding of the call-out information with no ambiguity, then the check-back can get a score of 3. But if there's only an acknowledgment like "Okay" or "Alright" but no major information that could clear out the ambiguity, that check-back utterance can only receive a score of 1. If the check-out phase contains some part of the key information in the call-out phase but has some level of ambiguity, the check-back utterance can get a score of 2. Table 26 provides the rubric and example for evaluating the check-back score.

M.3 Example Cases

CLC Phase	Semantic Types
Call-out	request, question, action directive, instruction, commitment, assert, knowledge sharing
Check-back	[another player] acknowledgment, confirm, (key information in call-out)
Closing-of-the-loop	[call-out speaker] acknowledgment, confirm, gratitude

Table 25: Common semantic types of CLC phases

Criteria	Example	Score
No confirmation of understanding	*Okay.*	1
Partial confirmation of understanding	*Okay, I am on my way.*	2
Full confirmation of understanding (key information repeated)	*Okay, I am on my way to B4 to clear the rubble.*	3

Table 26: Rubric for evaluating checkbacks in closed-loop communication events. The middle column shows examples of replies to the hypothetical call-out: *"Engineer, can you clear the rubble room B4?"*

Role	Utterance	CLC_Label	Checkback_Score
Green	where's the management meeting and the transporter here I'm going to go check in there	15a.16a	
Blue	okay	16b	1

Table 27: One sentence contains two events

Role	Utterance	CLC_Label	Checkback_Score
Red	transporter you at M1	42a	
Red	this is medic	42a+	
Green	this is transporter I am almost there	42b	2

Table 28: One sentence is cut off into several utterances

Role	Utterance	CLC_Label	Checkback_Score
Red	okay so E5 we should also be good	7a	
Blue	okay	7b	3
Blue	E5 looks good	7b+	3

Table 29: Two check-backs from one person for the same call-out. The scores should be the same for all "7b" labels because they are considered as one 7b event

Role	Utterance	CLC_Label	Checkback_Score
Red	yeah um can someone come with me to B2	30a	
Green	I'll be back there in a sec	30b-1	2
Blue	B2 yeah	30b-2	2

Table 30: Two check-backs for one call-out

Role	Utterance	CLC_Label	Checkback_Score
Red	I'm heading to A2 medic	12a	
Red	management meeting is in M3	13a	
Blue	B2 okay	12b.13b	1

Table 31: One check-back for two call-outs

Role	Utterance	CLC_Label	Checkback_Score
Green	this is transporter area c as in the hole is there a number associated or am I missing something	13a	
Blue	this is engineer I'm sorry I could not hear what you said could you repeat that for me please	13b	3
Green	B2 this is transporter you said that area C has Rubble	13c	
Green	oh Zone c i see	14a	
Blue	B2 yes on the south Zone C where the critical conditioner it got covered in rubble so I cleared it out I apologize	14b	3

Table 32: Follow-up questions for the call-out. The follow-up question is considered as a 3 scored *b*