Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2517–2527
Marseille, 20-25 June 2022

RoomReader: A Multimodal Corpus of Online Multiparty Conversational Interactions

Justine Reverdy,[*][†] Sam O'Connor Russell,[*][†] Louise Duquenne,[*]
Diego Garaialde,[‡] Benjamin Cowan,[‡] Naomi Harte[*][†]
[*] Sigmedia Group, ADAPT Centre, School of Engineering, Trinity College Dublin
[‡]ADAPT Centre, School of Information and Communication Studies, University College Dublin
[†]{reverdyj, russelsa, nharte}@tcd.ie

Abstract

We present RoomReader, a corpus of multimodal, multiparty conversational interactions in which participants followed a collaborative student-tutor scenario designed to elicit spontaneous speech. The corpus was developed within the wider RoomReader Project to explore multimodal cues of conversational engagement and behavioural aspects of collaborative interaction in online environments. However, the corpus can be used to study a wide range of phenomena in online multimodal interaction. The publicly-shared corpus consists of over 8 hours of video and audio recordings from 118 participants in 30 gender-balanced sessions, in the "in-the-wild" online environment of Zoom. The recordings have been edited, synchronised, and fully transcribed. Student participants have been continuously annotated for engagement with a novel continuous scale. We provide questionnaires measuring engagement and group cohesion collected from the annotators, tutors and participants themselves. We also make a range of accompanying data available such as personality tests and behavioural assessments. The dataset and accompanying psychometrics present a rich resource enabling the exploration of a range of downstream tasks across diverse fields including linguistics and artificial intelligence. This could include the automatic detection of student engagement, analysis of group interaction and collaboration in online conversation, and the analysis of conversational behaviours in an online setting.

Keywords: multimodal corpus, multiparty interaction, online interaction, video-conferencing, engagement detection

1. Introduction

To enable researchers to have a better understanding of complex communication behaviours that occur during social interactions, large labelled datasets of specific types of interactional situations are required (Ruhi et al., 2014). These datasets help uncover the links that tie high level phenomena such as conversational engagement, dominance or mutual understanding, and timed social signals contained in verbal and non-verbal behaviours (Anderson et al., 1991; Janin et al., 2003; McCowan et al., 2005; McKeown et al., 2011; Ringeval et al., 2013; Gupta et al., 2016). These links are explored to detect markers and find patterns that can be analysed and automatically extracted from features in audio and visual signals. Throughout 2020 and 2021, multiple lockdowns were enforced in many countries to slow the spread of the COVID-19 pandemic. Almost overnight online education became commonplace, previously having been a relatively niche practice. Workers, teachers and students, had to rapidly adapt to an unfamiliar paradigm to continue their professional, educational and social activities. This situation confronted many with the range of issues that accompany video-conferencing interactions, e.g., fatigue (Fauville et al., 2021), time latency leading to missed or distorted turn-taking cues (Seuren et al., 2021), or difficulties to perceive disengagement cues from interlocutors (Maimaiti et al., 2021). The situation also highlighted the lack of available datasets to allow researchers to study conversational engagement in online scenarios.

Thus in this paper we present RoomReader, a corpus of online, multiparty, multimodal interactions. The dataset is built around a University tutorial scenario, as the initial focus of our project was student engagement in online tutorials. However, the dataset also offers the prospect of studying multiparty online conversational interaction in a wider context. To the best of our knowledge, this is the first dataset dedicated to small group, human-human tutor-led online interactions focused on a conversational task. We provide 8 hours of high-quality multimodal recordings of conversational group interaction over Zoom, with rich annotations including a full transcription with utterance, word and phoneme level boundaries and labels for diverse phenomena such as engagement (self-reported and externally annotated) and paralinguistic elements (laughter, cough, etc.). The corpus provides researchers across a diverse range of fields, from linguistics to artificial intelligence, with a valuable resource to study how we interact and collaborate in multimodal online video-conferencing environments. Our second primary contribution is the adaptation of a continuous engagement annotation framework for online interactions. This enables the provision of continuous labels for student engagement. When combined with the perceptual self-reports of engagement which we collected from both students and tutors, this represents an exciting opportunity to explore the relationship between automatic engagement detection based on machine learning techniques, and self-reported psychometrics.

The corpus is also a rich resource for the wider exploration of mulitmodal aspects of conversational interaction online. The remainder of the paper is organised as follows: Section 2 reviews related work in the area of engagement detection and the wider context of online conversational interaction. Section 3 explains the scenario design while Section 4 describes the data collection, followed by the audio and video post-processing in Section 5. The transcriptions are detailed in Section 6, then the continuous engagement annotation process is outlined in Section 7, while Section 8 details the availability of the corpus.

2. Related work

In this section, we provide an overview of the literature on multimodal online interaction and engagement detection and hence outline the need for the a general-purpose, multimodal corpus annotated for engagement.

Human conversation is a multimodal process: when we interact, either in face-to-face or in video-conferencing settings, we produce and perceive verbal and non-verbal language in a continuous manner. Gestures such as head nods, for example, are not produced at random but rather form an integral part of the communication (McClave, 2000; Sundberg Cerrato, 2007), serving a variety of functions such as the facilitation of successful conversation through smooth turn-taking (Knapp et al., 2013). The multimodal aspects of human conversation have been extensively studied (Partan and Marler, 2005; D'Mello and Kory, 2015), through the analysis of annotated corpora.

Online video-conferencing platforms aim to provide an online virtual setting in which users can interact in a manner that mimics an in-person interaction. However, there are differences caused by the medium which can have negative effects on the user experience (Taylor, 2011; Hayakawa et al., 2017). The latency of online video-conferencing has been shown to have negative effects on the flow of conversation: through the analysis of online interactions in a medical setting, Seuren et al. show that the latency of online video-conferencing leads to difficulty in maintaining fluid conversation through interruptions or silences due to interlocutors not knowing when to speak (Seuren et al., 2021). Fatigue has also been associated with video-conferencing usage. Bailenson hypothesises that we knowingly compensate for the non-verbal communication deficits of online communication (e.g. latency, distorted eye gaze) by exaggerating our nonverbal body language gestures, which is a contributory factor in the widely-reported phenomenon of "Zoom fatigue" (Bailenson, 2021). Recent studies, of online video-conferencing users, shows that the level of fatigue experienced by the users varies with gender and ethnicity (Fauville et al., 2021; Ratan et al., 2021), highlighting equality issues in video-conferencing.

Despite some of the limitations mentioned above, video-conferencing enabled a form of interaction crucial to students all over the world to continue their education during the spread of the COVID-19 pandemic: remote synchronous courses.

While learning can happen in multiple forms, two types of phases can be distinguished in synchronous courses: listening phases that are typical of lectures, and conversational phases, that are more typical of tutorials. These two phases can be clearly divided within the structure of a course (e.g. a timetabled tutorial) or conversational episodes might happen within listening phases. The reason to distinguish these two phases is that they involve different mechanisms to display and perceive attention from teachers and students. The role of the student-tutor relationship is an often overlooked factor in student engagement (Farr-Wharton et al., 2018). It has been previously pointed out that universities with a learner-centred model had higher than average retention rates (Yorke and Thomas, 2003). Student-centred approaches that enhance student engagement are expected to be implemented by teaching staff that are already required to perform a large number of task within their teaching activities. For example, during tutorials, a high cognitive load is required, as attention is distributed between course/task material, handling conversation with students and monitoring student-student conversation, a context that is made even more difficult in online settings. A solution to alleviate the high pressure that is put on teachers and tutors is the development of automatic tools to assess students affects (i.e., interested, bored, confused, etc.) and engagement during video-conference sessions.

Engagement detection is a complex task (D'Mello et al., 2017) that is made all the more difficult in light of the fact that the concept of student engagement itself not being entirely agreed upon (Doherty and Doherty, 2018), but rather given different definitions. However, within these multiple definitions, it is widely acknowledged that engagement is a multi-dimensional meta-construct that can be divided into behavioural, emotional, cognitive and agentic engagement (Fredricks et al., 2004; Fredricks et al., 2016; Sinatra et al., 2015). Multimodal corpora for student engagement detection are still scarce (D'Mello, 2021), which slows down efforts made in automatic detection of engagement modelling that could help teachers and tutors have a better assessment of student engagement during courses and tutorials.

3. Scenario design

Student engagement is an important component in education as it has been repeatedly found to be a key factor in students' academic success and intention to persist in education (Schaufeli et al., 2002). Engagement is displayed differently during a traditional lecture as opposed to a conversation, where a number of social constraints related to speech interactions are relevant, e.g., turn-taking and verbal and non-verbal feedback. Conversations are highly social situations where participants may, for example, attempt to disguise negative emotions as a result of social pressure (Ekman and Friesen, 1969). Conversational engagement in an educational context can be defined as the degree of involvement of students in a topic being discussed and their willingness to continue the interaction. It can be analysed along three dimensions: from visual cues, from linguistic and paralinguistic cues, and from elements of the dialogue structure relevant to group cohesion.

Thus we wanted a scenario that would elicit spontaneous conversation and collaboration that was collectable during the lockdown phase of the COVID-19 pandemic. After reviewing a range of existing task-based interactions e.g., MapTask (Anderson et al., 1991), a decision was made to opt for a task that would not require written material or specific domain knowledge from participants.

We resolved on an adaptation of the MULTISIMO task (Koutsombogera and Vogel, 2018). Each session is led by an experienced University tutor. Three to five questions inspired from the TV show Family Feud are asked to a small group of students, requiring them to guess the three most popular answers to general knowledge questions that have been previously asked to a hundred people. Once the participants correctly guess all three answers, they are subsequently tasked with ranking the answers in order of their popularity. The students, through discussion, collaborate to establish the correct answers and the tutor guides them in a manner typical of a University tutorial, asking them for example to explain their reasoning, or telling them they are on the right track, or inviting others in the group to contribute. The tutors were fully informed about the aim of the task to elicit conversation and collaboration, and took part in a trial-run session (not included in the corpus) with feedback from the lead researcher. A sample of the tutor's dialogue illustrating the format of the scenario is provided below:

> *"I'm going to ask you a question that was also asked to 100 people in a survey and you guys are going to have to talk together and come up with the top three most popular answers the question. Name something people are often chased by in movies."*

It should be noted that whilst the primary focus of our project was student engagement in online tutorials, the scenario, as designed, allows a broader study of online multiparty multimodal conversation as the scenario elicits spontaneous dialogue and collaborative behaviours amongst participants.

4. Data collection

The data for the RoomReader corpus was collected at Trinity College Dublin (TCD), Ireland, between April and June 2021, for which 118 participants were recruited through an email campaign and personal connections, all being students or recent graduates of TCD. The two tutors, one male and one female, were engineering graduates with extensive experience of leading University tutorials at TCD. All data was captured and recorded on the TCD campus, however in order to reflect the real-world remote working and learning conditions adopted during the COVID-19 pandemic, there were no geographic restrictions imposed on participants. Participants were instructed to join on a computer device[1] in a location where they would consider appropriate for online remote learning. This led to the inclusion of a wide range of venues e.g., bedrooms, shared kitchens, and so on.

4.1. Ethical considerations

Ethical and privacy aspects were considered at all stages and integrated into the design process. The participant recruitment strategy, information provided to participants regarding the usage of their data, participant consent, data storage, and a licence agreement enabling the corpus to be shared with researchers worldwide, have been independently assessed by the School of Engineering Ethical Committee, TCD, and the TCD Data Protection Officer to ensure GDPR compliance.[2]

4.2. Participant information and consent

Each participant was given information about the purpose of the research project along with information on the recording process and the nature of the retained personal data such as age, gender, and country. Each participant was asked to fill out a consent form (provided in full in Appendix A) clearly indicating the usage of collected data. Once the collection was complete, all captured data (text and audio-video recordings) were checked to make sure no personal information was disclosed. An anonymisation process was conducted whereby any remaining personal information was removed. Participants could opt in or out of having their image shared in academic publications and video displayed at conferences. Participants could also opt-out of having their personality tests shared. Students were emailed a €15 gift card after the session.

4.3. Recording set-up

Each session was conducted using the video-conferencing platform Zoom (Zoom Video Communications Inc., 2019). Each session took the following general format: a session host, also being the researcher recording the session, would greet the participants onto the Zoom call and engage into a short warm up phase intended to break the ice between participants;[3] the participants were then given

[1] There are instances in the corpus where this was not followed and participants connected on mobile devices.

[2] European Union General Data Protection Regulation (GDPR) `https://gdpr-info.eu/` Last Accessed: 11.01.2022

[3] A series of video-conference ice-breakers were used, such as name changing to current mood or indicating one's level of energy for the day using a raised arm.

a random pseudonym to protect their real names; and then the recording would start whereby the host (named RoomReader) would mute and turn off their camera. From there, the session was led by the Tutor.

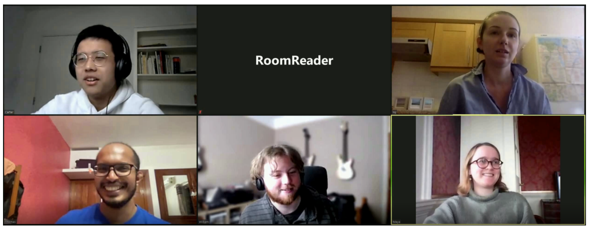


Figure 1: A still from the RoomReader corpus from session S09 captured with OBS Studio.

4.4. Recording outputs

Each session was recorded using two tools, the Zoom platform itself and an external third-party tool, OBS Studio.[4] Zoom comes with an in-built recorder which is turned on by the session host, capturing audio and video of the session. OBS Studio is a screen recording utility activated by the session host on their PC which captures the host's screen and PC audio. A combination of the two tools was used to ensure that there was robust recording in the event of a technology failure or human-error. The recording process generated a number of media files, each now described in this section. An outline of the files is provided in Table 4.

4.4.1. Audio

OBS Studio captures the system level audio and video from the host's PC, providing one MP4 video file per session. That file hence is a recording of the host's PC screen running the Zoom application, with audio of the conversation between participants in the session. The audio is ultra-wideband stereo with a 48kHz sample rate encoded with the AAC audio coding protocol. The audio is separated from the video and converted to a WAV file format using `ffmpeg`[5] for integration with tools such as Praat (Boersma and Van Heuven, 2001) without loss of quality or re-encoding. Zoom captures several 32kHz ultra-wideband mono audio files: an audio file containing all participants' speech (one per session) and individual audio tracks containing each speaker's isolated audio (one per speaker). Each file is an M4A file encoded with the AAC protocol and is converted to WAV audio. A full technical specification for the audio outputs for each session is provided in Table 1.

A number of audio processing algorithms[6] are applied by Zoom to reduce unwanted audio artefacts, including noise reduction and acoustic echo cancellation. We used the default Zoom audio processing configuration at the time of corpus collection in order to faithfully represent real-world conditions on the platform. We did not modify any settings in OBS Studio. Note that although it has higher bitrate than the Zoom audio (160kb/s and 126kb/s respectively), we found that the OBS Studio audio was overall lower in quality than the Zoom audio due to occasional clipping of the OBS Studio audio, resulting in the classic distortion of 'tinny' audio. Regardless, the Zoom audio is a high-quality recording.

Table 1: Audio capture from session recordings.

	Recording Utility	
Specification	**Zoom**	**OBS Studio**
Audio codec	AAC (LC)	AAC (LC)
Container	MP4A	MP4A
Sample rate	48 kHz	32 kHz
Bitrate	160 kb/s	126 kb/s

4.4.2. Video

OBS Studio captures quad high-definition (quad HD) 1440p video with a resolution of 2560x1440 pixels and a frame rate of 60 frames per second using the H.264 coding format in an MP4 file. Zoom captures standard high-definition (HD) 720p video with a resolution of 1280x720 pixels and a frame rate of 25 frames per second also using the H.264 format and an MP4 file extension. The OBS Studio video is of higher quality than the Zoom video upon inspection, as would be expected from higher resolution or number of pixels, and substantially higher frame rate. A full technical specification for the video files is provided in Table 2, and a still from the corpus illustrating the OBS Studio video recording is provided in Figure 1.

Table 2: Video capture from session recordings.

	Recording Utility	
Specification	**Zoom**	**OBS Studio**
Resolution	2560x1440 (quad HD)	1280x720 (standard HD)
Video codec	H.264	H.264
Container	mp4	mp4
Sample rate	60 fps	25 fps
Bitrate	15535 kb/s	609 kb/s

4.4.3. Automatic transcription

The in-built Zoom recording facility provided a basic transcription of the meeting. We chose to not use this transcription as it did not provide a sufficient level of detail to enable a linguistic analysis, lacking speaker labels, utterance and word-level time boundaries. We instead use a combination of and automatic speech recog-

[4] `https://obsproject.com/` Last Accessed: 11.01.2022

[5] `https://ffmpeg.org/` Last Accessed: 11.01.2022

[6] `https://support.zoom.us/hc/en-us/articles/360025379211-Zoom-Rooms-Audio-Guidelines` Last Accessed: 11.01.2022

nition (ASR) generated transcription and correction of the ASR transcription by a human expert. We detail this process in Section 6.

4.5. Metrics

Before the recording took place, each participant answered the Big Five personality test,[7] as it is acknowledged that personality traits have an influence on social behaviours that are integral to conversational group dynamic (John et al., 2008; Koutsombogera and Vogel, 2018). In the hours following the recordings, all students were asked to answer a short survey about their conversational engagement, and second survey relating to their perception of the tutor's behaviour during the session. Tutors were asked to answer a mirrored survey with the same questions in order to give their perception of conversational engagement for each student. The same survey was also given to the external annotators that rated the students according to the continuous scale described in Section 7.

Two additional surveys relating to group dynamic were also answered by the tutors for each session, after all sessions were complete. The purpose of these surveys was to measure their overall perception of group dynamic. The first group dynamic survey was specifically created for the corpus (referred to as the bespoke survey). The second group dynamic survey (referred to as the validated survey) was constructed using a previously-validated questionnaire, focusing on the assessment of group interaction quality in problem-based learning (PBL) (Visschers-Pleijers et al., 2005). We made slight adaptations to this survey to fit the online setting of our scenario. The tutors were given access to the recordings while completing these surveys, with freedom to watch the entire session if they required, to remind them of the session. For more details about the metrics, interested readers are referred to the documentation that will accompany the release (See Section 8).

Table 3 summarizes the surveys used to assess engagement (self-reported engagement from the students and students' engagement as perceived by tutors), group dynamic, personality tests and tutor's behaviour.

Table 3: Set of metrics, with respondent roles.

Measure	Role
Self-reported engagement	Student
Perceived engagement	Tutor + external annotators
Group dynamic (Bespoke)	Tutor
Group dynamic (Validated)	Tutor
Personality Test	Student & tutor
Tutor's behaviour	Student

4.6. Corpus description

We recorded 8h 44m across 30 sessions, with an average duration of 17m 29s. The sessions have been balanced for gender during recruitment, with half of the sessions being led by a female tutor and the other half led by a male tutor. Each session and each participant were assigned a unique identification number and a pseudonym, e.g. P045 'Max' is depicted in Figure 2. The recorded sessions contained a total of $n = 118$ participants (65 females, 51 males, 2 others, with 91 English L1, 27 English L2). When a participant indicated that they were not a native speaker of English, they were asked to self-declare English language fluency during the recruitment process. Participants who did not self-declare fluency were not recruited in order to ensure a high-level of English language proficiency across all participants. We recruited participants from a University setting, which is reflected in the younger age range of the participants (24 mean, 23 median, 43 max, 18 min).

Each session has been given two identifiers. One identifier starts with the letter 'S' followed by the number of the session in the order in which they were recorded (e.g., S01 to S30). A second identifier refers to a division of the sessions into 8 quads which are descriptors for the gender balancing element of the corpus design. We adapted this convention from the HCRC maptask corpus (Anderson et al., 1991). Each quad starts with the letter 'q', followed by the number of the quad (1 to 8), a letter indicating the gender of the tutor, then a letter indicating the gender of the participants (either x: mixed gender, f: female, m: male) and finally a number indicating the number of the session within the quad (1 to 4).[8]

Table 4: Overview of the files obtained from each recorded session.

	Recording Utility	
Media	Zoom	OBS Studio
Audio files	One audio track with all speakers and one audio track per speaker containing the speaker diaraised audio	One audio track with all speakers
Video files	One video file containing the recording of the session	One video file of the session captured from the host's computer
Miscellaneous	Automatic transcription from Zoom	

5. Post-processing

A number of post-processing steps were applied to the recordings obtained during corpus collection. These post-processing steps were necessary for several reasons: to avail of the highest quality audio and video

[7] The 44 items questionnaire assess five personality traits: Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness to Experience.

[8] For example, 'q1ff1' refers to the first session of the first quad with a female tutor and only female participants.

streams; to generate release versions of the corpus maximally useful for downstream linguistic and computational analyses; and to preserve the anonymity of participants where this was inadvertently compromised during the recording process. We outline these steps in this section.

To obtain the best quality audio and video for release, we combined the audio from Zoom with the video from OBS Studio. We found that the video from OBS Studio was superior in quality to that from Zoom (Section 4.4.2), but that the audio from Zoom was superior to that from OBS Studio (Section 4.4.1). The highest quality version of the recordings is therefore a version combining audio from Zoom and video from OBS Studio, and this is the version available to the research community.

A synchronisation process was necessary to combine the two modalities: having been captured from two independent tools, the audio from OBS Studio and the video from Zoom were not synchronised. There is a delay due to the time taken to initiate the Zoom recording and to launch OBS Studio. However, OBS Studio has an audio track which is synchronised with its video. We therefore indirectly synchronise the audio from Zoom with the video from OBS Studio by synchronisation of Zoom audio and OBS Studio audio.

To synchronise the audio tracks, we used cross-correlation, a measure of similarity between two time-varying signals (such as audio). The time delay is found by maximising the cross-correlation of the audio tracks, making use of the fact that by definition the cross-correlation of two signals is maximised when the signals are time-aligned (Schafer and Rabiner, 1975). We use the `correlate` function of the `scipy.signal`[9] Python package to compute the cross-correlation. A qualitative inspection of the synchronised audio files reveals that the method is highly accurate.

The original video recordings of the corpus contain all the participants in each session, arranged on a grid (Figure 1). In order to provide researchers with several versions of the corpus for different research tasks, we cropped the videos of each session to generate additional individual videos of each participant in a session (Figure 2). We used the OpenCV[10] video processing library in Python, and the resulting resolution of the videos was 834x472, which is the maximum resolution which can be obtained when cropping the participants in the recording of the Zoom grid. OpenCV was chosen over more powerful video editing tools such as Adobe Premiere Pro as it enabled the video to be cropped without compromising quality through re-encoding or compression. Two versions of each cropped individual participant video were generated. The first has the audio track with the session-level audio (i.e. all participants), and the second has only the audio from the participant depicted in the cropped video.

A number of videos required further editing. The participants' locations on the grid (Figure 1) was determined by Zoom, however their positions could change if a participant dropped off the call and had to rejoin it (e.g., due to a short interruption in internet connectivity). There were some instances of this behaviour in the recordings, necessitating further editing. There is no available video during the time in which a participant is missing from the call, so we filled in this time with empty (black) frames. Other editing steps pertained to anonymisation, such as placing in the participant's pseudonym where it had been replaced with their real name by Zoom upon rejoining a call after lost connectivity.

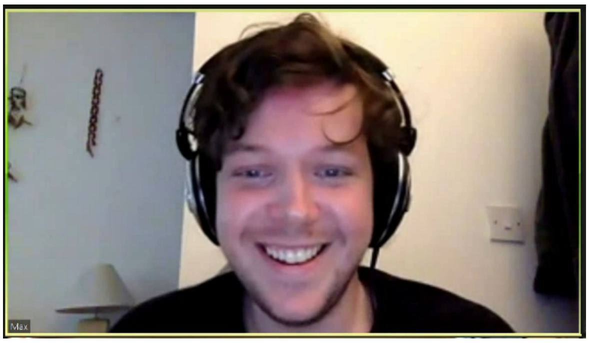


Figure 2: A still from a cropped video depicting participant P045 with pseudonym Max from session S06.

6. Transcription

Due to issues with the Zoom ASR generated transcription (Subsection 4.4.3), we conducted a semi-automated process to generate a more accurate and detailed transcription from scratch. Subsequent forced alignment (McAuliffe et al., 2017) yielded word and phoneme level boundaries. Transcription files are provided in the TextGrid format for ease of use with the Praat (Boersma and Van Heuven, 2001) annotation and linguistic analysis software. The transcription generation and correction process is outlined in this section.

6.1. Automatic speech recognition

We used the cloud-based speech-to-text service from Microsoft Azure[11] to obtain an initial transcription which we later manually corrected. The tool provides a number of English language models including United States English (EN-US) and United Kingdom English (EN-UK).[12] Our recordings mostly contained Irish native speakers of English, with Irish-English a recog-

[9] `https://docs.scipy.org/doc/scipy/reference/signal.html` Last Accessed: 11.01.2022

[10] `https://opencv.org/` Last Accessed: 11.01.2022

[11] `https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/speech-to-text` Last Accessed: 11.01.2022

[12] An Irish-English model was unavailable at the time so we do not present a WER for this model. It has since been added by Microsoft to their speech-to-text service.

nised distinct dialect of English. We conducted an evaluation of the EN-US and EN-UK ASR models on a subset of hand-transcribed recordings containing four native speakers and one non-native speaker of English. We used the Microsoft ASR tool with both EN-US and EN-UK models. We found that the EN-US model was more accurate than the EN-UK model with WER of 25.1% and 30.1% respectively, averaged across all the speakers. For the four native speakers of English, the WER was 13.8% and 10.5% respectively. We have not systematically investigated this result, though there are notable similarities between the dialects of Standard US English and most Irish-English dialects not commonly present in many dialects of English spoken in the UK, such as rhoticity (O'Sullivan, 2013; Demirezen, 2012). We note the reduced accuracy of the ASR on non-native speakers of English, and this was a factor in our decision to correct the transcriptions using an expert annotator (Section 6.2).

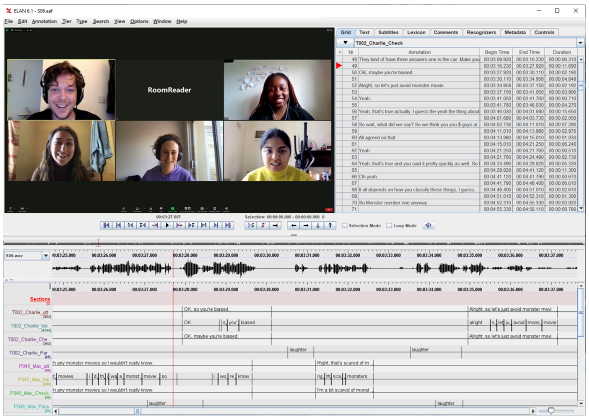
We obtained automatic transcriptions for the Zoom audio of each session using the Microsoft EN-US ASR model. Each session had one audio file per participant containing their dialogue resulting in one transcription per participant. To obtain a full transcription of the session with speaker labels, we joined these transcriptions into a single TextGrid file. Synchronisation was not an issue as all audio tracks are synchronised automatically by Zoom. We enabled utterance and word-level time boundaries in the settings of the ASR tool for each audio track. A complimentary set of word boundaries and additional phoneme boundaries was obtained using the Montreal Forced Aligner (McAuliffe et al., 2017) with the provided model and pronunciation dictionary derived from the LibriSpeech corpus (Panayotov et al., 2015).

6.2. Manual correction of transcription and paralinguistic elements

The transcriptions have been manually corrected from the automatic transcripts to ensure the high quality and accuracy necessary to investigate linguistic phenomena. Each utterance and utterance boundary from the Microsoft ASR tool has been inspected using PRAAT software (Boersma and Van Heuven, 2001). Utterance boundaries were subsequently corrected if deemed necessary, and the orthographic format of Microsoft ASR was respected but corrected in case of errors. The Microsoft ASR tool did not consistently record disfluencies in the dialogues. We therefore reintroduced any disfluencies that were eliminated by the automatic transcription process during manual correction (i.e. restarts, repetitions or repairs) to correspond to naturally occurring speech, with usage of the Switchboard Transcription Guidelines 7.1 for speech transcriptions (Hamaker et al., 1998) for elements outside Microsoft orthographic format.

An additional layer of manual annotation added to the corpus relates to the paralinguistic elements that both surround and are an integral part of naturally occurring speech, e.g. laughter (which represents the majority of the annotated paralinguistic elements in the corpus). We annotated the corpus for a subset of paralinguistic elements chosen from the list of typical non-speech sounds in the Switchboard Transcription Guidelines (Hamaker et al., 1998).

Following manual corrections and paralinguistic phenomenon annotation, the files were imported into the ELAN software (Wittenburg et al., 2006). ELAN was chosen as it enables the complete hierarchy of annotations (utterances, words, parlinguistics) to be simultaneously visualised. The ELAN tool with corpus annotations is shown in Figure 3). With ELAN, the scenario have been be divided into sections and subsections corresponding to the question and ranking phases from the designed scenario, and the format can facilitates the addition of further annotations in the future.



T002_Charlie_Check

>	Nr	Annotation	Begin Time	End Time	Duration
	48	They kind of have three answers one is the car. Make you	00:03:09.920	00:03:16.230	00:00:06.310
▶	49		00:03:16.230	00:03:27.920	00:00:11.690
	50	OK, maybe you're biased.	00:03:27.920	00:03:30.110	00:00:02.190
	51		00:03:30.110	00:03:34.958	00:00:04.848
	52	Alright, so let's just avoid monster movie.	00:03:34.958	00:03:37.150	00:00:02.192
	53		00:03:37.150	00:03:41.050	00:00:03.900
	54	Yeah.	00:03:41.050	00:03:41.760	00:00:00.710
	55		00:03:41.760	00:03:46.030	00:00:04.270
	56	Yeah, that's true actually. I guess the yeah the thing about	00:03:46.030	00:04:01.680	00:00:15.650
	57		00:04:01.680	00:04:03.730	00:00:02.050
	58	So wait, what did we say? So we think you you $ guys ar.	00:04:03.730	00:04:11.010	00:00:07.280
	59		00:04:11.010	00:04:13.980	00:00:02.970
	60	All agreed on that.	00:04:13.980	00:04:15.010	00:00:01.030
	61		00:04:15.010	00:04:21.250	00:00:06.240
	62	Yeah.	00:04:21.250	00:04:21.760	00:00:00.510
	63		00:04:21.760	00:04:24.490	00:00:02.730
	64	Yeah, that's true and you said it pretty quickly as well. So I	00:04:24.490	00:04:29.820	00:00:05.330
	65		00:04:29.820	00:04:41.120	00:00:11.300
	66	Oh yeah.	00:04:41.120	00:04:41.790	00:00:00.670
	67		00:04:41.790	00:04:48.400	00:00:06.610
	68	It all depends on how you classify these things, I guess.	00:04:48.400	00:04:51.010	00:00:02.610
	69		00:04:51.010	00:04:52.310	00:00:01.300
	70	So Monster number one anyway.	00:04:52.310	00:04:55.330	00:00:03.020
	71		00:04:55.330	00:04:56.110	00:00:00.780

00:03:27.697 Selection: 00:00:00.000 - 00:00:00.000 0

Sections [0]
T002_Charlie_utt [364]: OK, so you're biased. | Alright, so let's just avoid monster movi
T002_Charlie_tok [4346]: OK | s | you' | biased | alright | s | let | ju | avoid | mons | movie
T002_Charlie_Che [362]: OK, maybe you're biased. | Alright, so let's just avoid monster movi
T002_Charlie_Par [65]: laughter | laughter
P045_Max_utt [96]: h any monster movies so I wouldn't really know. | Right, that's scared of m
P045_Max_tok [636]: t | movies | i | d | th | i | wa | a | monst | movie | so | i | wo | re | know | rig | th | sca | monsters
P045_Max_Check [98]: h any monster movies so I wouldn't really know. | I'm a bit scared of monst
P045_Max_Para [39]: laughter | laughter

Figure 3: Example of a session in ELAN with participants transcription tiers (ASR and manually corrected).

7. Data annotation

The manual annotation of students' behaviour was made using an instrument created originally for classroom settings by Goldberg et al. (Goldberg et al., 2019; Sümer et al., 2021), that we adapted to video-conferencing group tutorials by shifting the focus from full-body language to rather focusing on facial expressions and attention. The adapted scale is shown in Figure 4. The scale follows a combination of off-task/on-task and passive/interactive concepts, and ranges from +2 for high observed engagement to -2 for active disruption, while 0 indicates a passive behaviour.

The continuous engagement annotations of student participants were made using the CARMA software (Girard, 2014) as shown in Figure 5 by two expert annotators recruited at Trinity College Dublin, specifically trained to follow the above scale. Each annotator rated half of the sessions using the public release videos of the corpus. We provided the annotators with videos of individual participants which were paired with an audio track containing dialogue from all speakers in the session (see Section 5 for further detail). As such, an

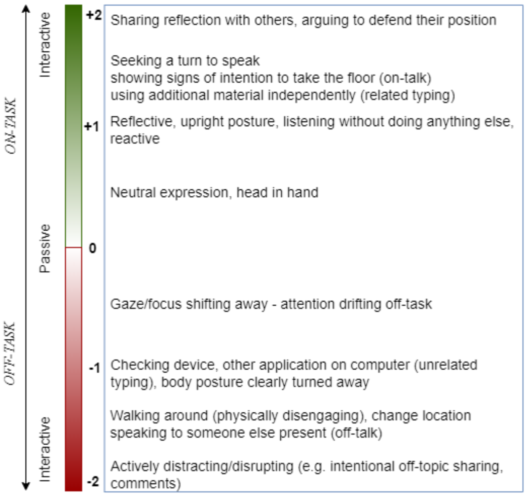

ON-TASK
OFF-TASK
Interactive
Passive
Interactive
+2
+1
0
-1
-2
Sharing reflection with others, arguing to defend their position
Seeking a turn to speak
showing signs of intention to take the floor (on-talk)
using additional material independently (related typing)
Reflective, upright posture, listening without doing anything else, reactive
Neutral expression, head in hand
Gaze/focus shifting away - attention drifting off-task
Checking device, other application on computer (unrelated typing), body posture clearly turned away
Walking around (physically disengaging), change location speaking to someone else present (off-talk)
Actively distracting/disrupting (e.g. intentional off-topic sharing, comments)

Figure 4: Adapted continuous scale for online conversational engagement in groups.

annotator receives the benefit of hearing all participants in the session but focusing visually on that a single student. The order in which the individual videos of each session were annotated was randomised in order to reduce repetitiveness of the task, as excessive task repetition could fatigue the annotator, reducing the quality and accuracy of the annotations.

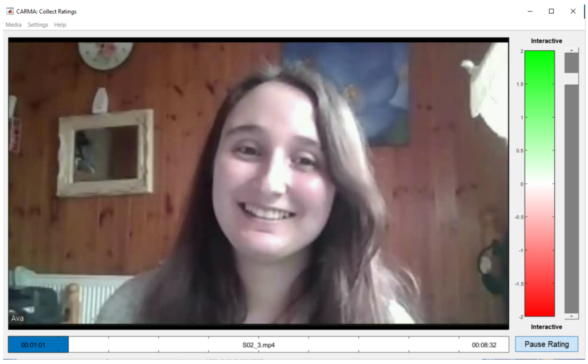


Figure 5: Example of continuous engagement rating using the CARMA software.

One issue that requires careful attention in annotations of high level behavioural phenomena such as engagement (Inoue et al., 2019) is the potential subjectivity that arises as each defined state requires interpretation. Previous studies investigating engagement used majority voting and others assumed that a latent character is affecting each annotator's perception (Inoue et al., 2019). We make individual annotations available to allow assessment of inter-rater agreement.

For completeness, a number of session level descriptors are released for each session under the following categories, with sample descriptors provided in brackets: screen interventions (blanking participants home screen for privacy), video events, video description, video lighting (uneven lighting), video background, participant position in frame (occlusions), facial features (beard, glasses), online accessories (headphones).

8. RoomReader corpus availability

From its inception, this dataset has been planned as a structured corpus to be made available to researchers across the wide range of fields interested in speech interaction analysis.[13] A summary of final audio, video, annotations that will be released to the wider research community can be seen in Table 5.

Table 5: Summary of content of corpus release.

Category	Description
Audio	Participant-level audio, one track per participant Session-level audio, one track per session
Video	Video of every participant on the call, arranged in a grid, one per session A cropped video of each participant in the grid, one per participant
Transcription	ASR-generated transcription of session dialogue with speaker labels, TextGrid format Rich manually-corrected transcription of session dialogue with speaker labels, TextGrid format Utterance, word and phoneme-level boundaries for each session, TextGrid
Annotation	Continuous annotation of student engagement (4 external annotators) Paralinguistic annotations (e.g., laughter, coughs, etc.) Session-level observations (e.g., uneven lighting, wearing headphones)
Psychometrics	Participant personality tests, where consent provided Participant demographic information (age, gender, country) Students' self-reported levels of engagement, Likert scale Tutors' perception of student levels of engagement Tutor behaviour as rated by students, Likert scale

9. Conclusion

This paper has presented the RoomReader corpus, a labelled dataset of 30 online tutorial-style sessions involving 118 participants. The audio and video material is accompanied by automatically and manually corrected transcriptions as well as a rich set of associated engagement annotations, group cohesion and additional information concerning participants such as personality tests. We are making this corpus freely-available for research under a non-commercial license. We believe this corpus is a substantial resource that will allow researchers study many aspects of multimodal, multiparty conversations in online settings.

10. Acknowledgements

This research was supported by Science Foundation Ireland under Grant Agreement No. 20/COV/8537 and 13/RC/2106_P2 at the ADAPT SFI Research Centre at Trinity College Dublin and University College Dublin.

[13] https://sigmedia.github.io/resources/datasets

11. Bibliographical References

Bailenson, J. N. (2021). Nonverbal overload: A theoretical argument for the causes of zoom fatigue. *Technology, Mind, and Behavior*, 2(1).

Boersma, P. and Van Heuven, V. (2001). Speak and unspeak with praat. *Glot International*, 5(9/10):341–347.

Demirezen, M. (2012). Which /r/ are you using as an english teacher? rhotic or non-rhotic? *Procedia-Social and Behavioral Sciences*, 46:2659–2663.

D'Mello, S. K. and Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36.

D'Mello, S. K., Dieterle, E., and Duckworth, A. (2017). Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational psychologist*, 52(2):104–123.

D'Mello, S. K. (2021). Improving student engagement in and with digital learning technologies. *OECD Digital Education Outlook 2021 Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots*, pages 79–104.

Doherty, K. and Doherty, G. (2018). Engagement in HCI: conception, theory and measurement. *ACM Computing Surveys (CSUR)*, 51(5):1–39.

Ekman, P. and Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica*, 1(1):49–98.

Farr-Wharton, B., Charles, M. B., Keast, R., Woolcott, G., and Chamberlain, D. (2018). Why lecturers still matter: the impact of lecturer-student exchange on student engagement and intention to leave university prematurely. *Higher Education*, 75(1):167–185.

Fauville, G., Luo, M., Muller Queiroz, A. C., Bailenson, J. N., and Hancock, J. (2021). Nonverbal mechanisms predict zoom fatigue and explain why women experience higher levels than men. *Available at SSRN 3820035*.

Fredricks, J. A., Blumenfeld, P. C., and Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1):59–109.

Fredricks, J. A., Filsecker, M., and Lawson, M. A. (2016). Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues. *Learning and Instruction*, 43:1–4.

Girard, J. M. (2014). Carma: Software for continuous affect rating and media annotation. *Journal of open research software*, 2(1).

Goldberg, P., Sümer, Ö., Stürmer, K., Wagner, W., Göllner, R., Gerjets, P., Kasneci, E., and Trautwein, U. (2019). Attentive or not? toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educational Psychology Review*, pages 1–23.

Hamaker, J., Zeng, Y., and Picone, J. (1998). Rules and guidelines for transcription and segmentation of the switchboard large vocabulary conversational speech recognition corpus. *Mississippi State University, Tech. Rep*.

Hayakawa, A., Vogel, C., Campbell, N., and Luz, S. (2017). Perception changes with and without a video channel: A study from a speech-to-speech, machine translation mediated map task. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000401–000406. IEEE.

Inoue, K., Lala, D., Takanashi, K., and Kawahara, T. (2019). Latent character model for engagement recognition based on multimodal behaviors. In *9th International Workshop on Spoken Dialogue System Technology*, pages 119–130. Springer.

John, O. P., Naumann, L. P., and Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. *Handbook of personality: Theory and research*, pages 114–158.

Knapp, M. L., Hall, J. A., and Horgan, T. G. (2013). *Nonverbal Communication in Human Interaction*. Boston, MA: Cengage Learning.

Koutsombogera, M. and Vogel, C. (2018). Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2945–2951.

Maimaiti, G., Jia, C., and Hew, K. F. (2021). Student disengagement in web-based videoconferencing supported online learning: an activity theory perspective. *Interactive Learning Environments*, pages 1–20.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of pragmatics*, 32(7):855–878.

O'Sullivan, J. (2013). Advanced dublin english in irish radio advertising. *World Englishes*, 32(3):358–376.

Partan, S. R. and Marler, P. (2005). Issues in the classification of multimodal communication signals. *The American Naturalist*, 166(2):231–245.

Ratan, R., Miller, D. B., and Bailenson, J. N. (2021). Facial appearance dissatisfaction explains differences in zoom fatigue. *Cyberpsychology, Behavior, and Social Networking*, 25(2):124–129.

Ruhi, Ş., Haugh, M., and Schmidt, T. (2014). *Best Practices for Spoken Corpora in Linguistic Research*. Cambridge Scholars Publishing.

Schafer, R. W. and Rabiner, L. R. (1975). Digital representations of speech signals. *Proceedings of the IEEE*, 63(4):147–148.

Schaufeli, W. B., Martinez, I. M., Pinto, A. M., Salanova, M., and Bakker, A. B. (2002). Burnout and engagement in university students: A cross-national study. *Journal of cross-cultural psychology*, 33(5):464–481.

Seuren, L. M., Wherton, J., Greenhalgh, T., and Shaw, S. E. (2021). Whose turn is it anyway? latency and the organization of turn-taking in video-mediated interaction. *Journal of pragmatics*, 172:63–78.

Sinatra, G. M., Heddy, B. C., and Lombardi, D. (2015). The Challenges of Defining and Measuring Student Engagement in Science. *Educational Psychologist*, 50(1):1–13.

Sümer, Ö., Goldberg, P., D'Mello, S., Gerjets, P., Trautwein, U., and Kasneci, E. (2021). Multimodal engagement analysis from facial videos in the classroom. *arXiv preprint arXiv:2101.04215*.

Sundberg Cerrato, L. (2007). *Investigating communicative feedback phenomena across languages and modalities*. Ph.D. thesis, KTH Royal Institute of Technology, Stockholm.

Taylor, T. (2011). Video conferencing us talking face-to-face: Is video suitable for supportive dialogue? *International Journal of Therapy and Rehabilitation*, 18(7):392–402.

Visschers-Pleijers, A. J., Dolmans, D. H., Wolfhagen, I. H., and van der Vleuten, C. P. (2005). Development and validation of a questionnaire to identify learning-oriented group interactions in pbl. *Medical teacher*, 27(4):375–381.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: A professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.

Yorke, M. and Thomas, L. (2003). Improving the retention of students from lower socio-economic groups. *Journal of higher education policy and management*, 25(1):63–74.

Zoom Video Communications Inc. (2019). Zoom meetings & Chat. Retrieved from `https://zoom.us/meetings`.

12. Language Resource References

Anderson, Anne H and Bader, Miles and Bard, Ellen Gurman and Boyle, Elizabeth and Doherty, Gwyneth and Garrod, Simon and Isard, Stephen and Kowtko, Jacqueline and McAllister, Jan and Miller, Jim and others. (1991). *The HCRC map task corpus*. SAGE Publications Sage UK: London, England.

Gupta, A., D'Cunha, A., Awasthi, K., and Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*.

Janin, Adam and Baron, Don and Edwards, Jane and Ellis, Dan and Gelbart, David and Morgan, Nelson and Peskin, Barbara and Pfau, Thilo and Shriberg, Elizabeth and Stolcke, Andreas and others. (2003). *The ICSI meeting corpus*.

Koutsombogera, Maria and Vogel, Carl. (2018). *Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus*.

McCowan, Iain and Carletta, Jean and Kraaij, Wessel and Ashby, Simone and Bourban, S and Flynn, M and Guillemot, M and Hain, Thomas and Kadlec, J and Karaiskos, Vasilis and others. (2005). *The AMI meeting corpus*.

McKeown, Gary and Valstar, Michel and Cowie, Roddy and Pantic, Maja and Schroder, Marc. (2011). *The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent*. IEEE.

Panayotov, Vassil and Chen, Guoguo and Povey, Daniel and Khudanpur, Sanjeev. (2015). *Librispeech: an asr corpus based on public domain audio books*.

Ringeval, Fabien and Sonderegger, Andreas and Sauer, Juergen and Lalanne, Denis. (2013). *Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions*.

A. Appendix: RoomReader Corpus Participant Consent Form

Table 6: Consent form given to the RoomReader corpus participants.

There are 16 sections in this form. Each section has a statement and asks you to tick the box if you agree. Please ask any questions you may have when reading each of the statements, by emailing us at roomreader@adaptcentre.ie. Please answer no if you do not agree. Thank you for participating.

1. I confirm I have read and understood the Participant Information Form for the study. The information has been fully explained to me and I have been able to ask questions, all of which have been answered to my satisfaction.
2. I understand that this study is entirely voluntary, and if I decide that I do not want to take part, I can withdraw my consent at any time for up to two weeks after the recording, without giving a reason.
3. I understand that I will be paid for taking part in this study.
4. I agree to take part in this research study having been fully informed of the risks, benefits and alternatives which are set out in full in the Information Form which I have been provided with.
5. I know how to contact the research team if I need to.
6. I agree to being contacted by researchers by email as part of this research study.
7. I agree to complete a questionnaire before and after the recording as part of this research study.
8. I agree that researchers worldwide, which will be bound to a license agreement to only use the data for research and learning purposes, can use stills from videos which may contain my face, in publishing or presenting research papers related to this work. (Ticking ”No” here does not exclude you from the study.)
9. I agree that researchers worldwide, which will be bound to a license agreement to only use the data for research and learning purposes, can publicly play videos which may contain my face and voice, in presenting research related to this work. (Ticking ”No” here does not exclude you from the study.)
10. I understand that any identifiable information about me (personal data) will be protected in accordance with the General Data Protection Regulation (GDPR).
11. I understand that I am participating in a study that will record my face and voice, and this means that I can potentially be identified.
12. I agree that the questionnaires may be shared with third party researchers worldwide for research and learning purposes. This concerns the post-recording self-assessment, and the assessment of the facilitator behaviour. I understand that the questionnaire will have a code number to replace my name.
13. I agree that the questionnaires may be shared with third party researchers worldwide for research and learning purposes. This concerns the personality test. I understand that the questionnaire will have a code number to replace my name. (Ticking ”No” here does not exclude you from the study.)
14. I understand that I have the right to review my recording within two weeks following the session.
15. I understand that partial recordings may be shared with third party academics worldwide for research and learning purposes.
16. I understand that the original recording of my session will be retained by Trinity College Dublin for 5 years for use solely by Trinity College Dublin, and then destroyed.