Interspeech 2024
1-5 September 2024, Kos, Greece

Understanding "understanding": Presenting a richly annotated multimodal corpus of dyadic interaction

Leonie Schade, Nico Dallmann, Olcay Türk, Petra Wagner

Bielefeld University, Germany
leonie.schade@uni-bielefeld.de

Abstract

This paper presents the MUNDEX corpus (MUltimodal UNDerstanding of EXplanations) together with past and current investigations using its data. The corpus is constructed to observe the dynamics of co-constructed communication and the understanding of explanations on multiple modalities in dyadic interactions. These modalities are annotated on several levels, including orthographic transcriptions, acoustic information, annotations of head movement, gaze, manual gestures, and further non-verbal behaviour as well as discourse annotations. Present and past projects are also concerned with adding further to these annotations. The interlocutors' level of understanding is currently investigated in regard to several verbal and non-verbal behaviour markers of both the explaining and listening side.

Index Terms: multimodality, discourse, understanding, explanation, feedback

1. Introduction

The direct or indirect signalling of having understood an interlocutor's message is a crucial aspect of everyday communication, and of particular relevance in interactive explanations. The MUNDEX corpus [1] is constructed with an interest in multimodal interaction dynamics connected to different levels of understanding within ongoing explanations. These dynamics are not only of relevance for investigations of human-human communication, but also for human-machine interaction. In communication, speakers and listeners make use of rich and complex combinations of verbal and nonverbal signals, taking into account situative and communicative needs as well as available resources (e.g., mutual visibility, level of environmental noise, relevance of message etc.). To this day, many facets to multimodal behaviour, including those connected to signalling or reacting to the signalling of (non-)understanding, are not well-understood [2]. This is largely due to their subtle nature, variability, and also the gradual quality of understanding itself [3]. The MUNDEX corpus aims to capture these phenomena and their lead-to and follow-up during on-line interactions. Therefore, it adopts a dyadic paradigm involving an explanation of a board game by a self-trained explainer (ER) to a naive explainee (EE), which are video-recorded from multiple perspectives to allow comprehensive monitoring of all related behaviours.

2. The Corpus

2.1. Recordings

88 dyads (one ER and one EE, not controlled for age or gender), all speakers of German, were recorded. Within the study design, each ER explained the game to at least two EEs (32 ERs, 88 EEs), which enables examining differences in individual explanation strategies and multimodal cues while the ER was held constant. Prior to recording, the ER had familiarized themselves with the game and its rules while the EE was a novice. Every recording began with the explanation of a strategic board game (Deep Sea Adventure), followed by gameplay. The ER's task was to explain the game without having the instructions or any game pieces visible to increase the information gap to potentially trigger a greater variety of different levels of understanding. Then, each dyad played up to three rounds of the game together. At the end, both the ER and the EE also were recorded during a *video-recall* – a thinking-aloud-task where they separately re-watched and commented on the recording of their previous interaction. The EE was asked to stop the video whenever they thought they had or had not understood the current explanation and comment on their state of understanding. The ERs were asked to do comment on their belief about the EE's level of understanding. This way, a detailed account of (beliefs about) various levels of understanding from different perspectives was captured. Each interaction lasted about 30 minutes. Including the recall tasks, 143 hours of recordings were collected in total. Every interaction was recorded with separate headset microphones for ER and EE, and with six cameras capturing the interaction at different angles (1920x1080, 50fps).

2.2. Annotations

2.2.1. Transcription and Acoustic Information

All recordings were automatically transcribed using the BAS Web-service [4] or OpenAI's Whisper [5]. They were then corrected manually using Praat [6], adding labels for disfluencies, backchannels, laughter, and breathing. The extraction of acoustic information on voice quality, intensity, and pitch dynamics as well as symbolic prosodic annotations [7] are planned.

2.2.2. Multimodal Annotations

Presently, ELAN annotations [8] of head movement, gaze, and manual gestures of 21 EEs exist for reports of understanding in the video-recalls, as well as randomly chosen control intervals. Annotations of the EE's head movements included *nods* and *shakes* [2]. Gaze was classified in three broad gaze positions: gaze towards the other *player*, towards the *game*, or *away*, with the annotations referring to the shifts between two positions. Manual gestures (*iconic*, *deictic*, *metaphoric*, and *beat gestures*) were annotated separately for each hand [9]. A tool for gaze annotation using OpenFace [10] was developed [11] and is set to be used. Other tools for head gesture annotation using [10] and for manual gesture annotation using [12] are in development. MUNDEX contains further annotations of non-verbal

behaviour, including torso movements, facial expressions, and adapters such as sitting adjustments and fidgeting [9]. Torso movements (i.e., leans) were annotated as shifts between *backward* and *forward* positions with additional annotations of *sideward* leans while also including shoulder shrugs. For the time being, there are binary annotations (absence/presence) of facial expressions [13]. As for adaptors, any adjustments and adaptations by the EE of either objects or their own body were annotated into one of the following six categories: (1) hand fidgeting, (2) adjustment of clothing, (3) self-touch, (4) adjustment of seating position, (5) relocation of game pieces, or (6) fidgeting with game pieces.

2.2.3. Discourse Annotation

Discourse annotations are based on modified versions of DAMSL [14] and DIT++ [15]: Annotations in MUNDEX are made in correspondence to three main categories: *forward looking*, *backward looking*, and *dialog control functions*. These exist within a hierarchy with their sub-categories as in the original schemes. The functions are not mutually exclusive — one utterance can have more than one main function. Here, we describe briefly only one of the categories, i.e., *backward looking functions*: *Backward looking functions* indicate how the current utterance relates to previous turns in the dialog, differentiated into answers, feedback utterances, and the communication of understanding. There are three categories of feedback utterances: *auto-feedback* applies to instances when a speaker gives information about their processing of the other's utterance, differing from *allo-feedback*, which can be conceived to be feedback to feedback. *Self-feedback* then refers to feedback on the speaker's own actions or utterances. *Backward looking functions* are at the core of the corpus since these contain annotations of understanding. An utterance can be classified as (1) a resolution of a misunderstanding, (2) a taking of the other's turn to demonstrate understanding, (3) a hold of the conversation to gain more time, (4) a correction of the other's utterance by replying with the correct answer, (5) a repetition and paraphrasing can be used to signal understanding, and/or (6) the speaker can signal they have corrected a self-error. Additionally, the level of understanding (understanding, non-understanding, misunderstanding, partial understanding) can be added.

3. Discussion

The data in the MUNDEX corpus has allowed investigations into listeners' gaze behaviour in dyadic interactions and whether it follows the entropy rate constancy principle [11]. One present project investigates whether an EE's level of understanding (as extracted from the video-recalls) can be reliably predicted via machine learning models making use of the multimodal cues annotated in the corpus. The use of deictic co-speech gestures of an interlocutor in regard to their dyadic counterpart's understanding is also examined by [16]. Furthermore, deictic gestures and their use in connection to discourse markers representing understanding are studied. The investigation of non-verbal signals, e.g., eye-contact, is planned for the near future. Another ongoing study investigates turn taking dynamics, systematically comparing phases of different levels of understanding and non-understanding, but also investigating whether explanations differ from other dyadic interactions.

4. Acknowledgements

This research was funded by the German Research Foundation (DFG): TRR 318/1 2021–438445824. The first two authors of the paper claim no ownership of the corpus. They were involved in the data collection and annotations, not the conceptual work on the corpus. All credit goes to the project members (https://trr318.uni-paderborn.de/en/projects/a02).

5. References

[1] O. Türk, P. Wagner, H. Buschmeier, A. Grimminger, Y. Wang, S. Lazarov, P. Paggio, and P. Prieto, "MUNDEX: A multimodal corpus for the study of the understanding of explanations," in *1st International Multimodal Communication Symposium. Book of Abstracts*, 2023.

[2] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Communication*, vol. 57, pp. 209–232, 2014.

[3] H. Buschmeier and S. Kopp, "Communicative listener feedback in human–agent interaction: Artificial speakers need to be attentive and adaptive," in *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, 2018.

[4] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, Sep. 2017.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[6] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." [Online]. Available: https://www.praat.org

[7] F. Kügler, S. Baumann, B. Andreeva, B. Braun, M. Grice, J. Neitsch, O. Niebuhr, J. Peters, C. Röhr, A. Schweitzer, and P. Wagner, "Annotation of german intonation: DIMA compared with other annotation systems," in *Proceedings of the 19th International Congress of Phonetic Sciences*, 2019, pp. 1297–1301.

[8] T. L. Archive, "Elan," 2023. [Online]. Available: https://archive.mpi.nl/tla/elan

[9] D. McNeill, *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press, 1992.

[10] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[11] Y. Wang and H. Buschmeier, "Does listener gaze in face-to-face interaction follow the entropy rate constancy principle: An empirical study," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, 2023, pp. 15 372–15 379. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.1026

[12] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for building perception pipelines," *ArXiv*, vol. abs/1906.08172, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:195069430

[13] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[14] M. G. Core and J. F. Allen, "Coding dialogs with the DAMSL annotation scheme," in *Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, 1997.

[15] H. Bunt, "The DIT++ taxanomy for functional dialogue markup," in *Proceedings of the AAMAS 2009 Workshop 'Towards a Standard Markup Language for Embodied Dialogue Acts'*, 2009, pp. 13–23.

[16] S. Lazarov and A. Grimminger, "Variations in explainers' gesture deixis in explanations related to the monitoring of explainees' understanding," in *Proceedings of the Annual Meeting of the Cognitive Science Society, 46*, 2024. (under review).