Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 743–749
Marseille, 11–16 May 2020

Chinese Whispers: A Multimodal Dataset for Embodied Language Grounding

Dimosthenis Kontogiorgos, Elena Sibirtseva, Joakim Gustafson
KTH Royal Institute of Technology
Stockholm, Sweden
{diko,elenasi,jkgu}@kth.se

Abstract

In this paper, we introduce a multimodal dataset in which subjects are instructing each other how to assemble IKEA furniture. Using the concept of 'Chinese Whispers', an old children's game, we employ a novel method to avoid implicit experimenter biases. We let subjects instruct each other on the nature of the task: the process of the furniture assembly. Uncertainty, hesitations, repairs and self-corrections are naturally introduced in the incremental process of establishing common ground. The corpus consists of 34 interactions, where each subject first assembles and then instructs. We collected speech, eye-gaze, pointing gestures, and object movements, as well as subjective interpretations of mutual understanding, collaboration and task recall. The corpus is of particular interest to researchers who are interested in multimodal signals in situated dialogue, especially in referential communication and the process of language grounding.

Keywords: multimodal interaction, situated dialogue and discourse, grounding, referential communication, non-verbal signals.

1. Introduction

Recently, a large body of research has focused in developing robots' communicative and social skills (Thomaz et al., 2016). In multiparty settings, robots are going to learn by observing what task humans perform and how they instruct each other using a combination of voice, gaze and gestures. Collaborative robots should through multimodal output be able to convey their perception and level of understanding of human actions, and they should be able assess when humans need further assistance.

Detecting users' intrinsic states of understanding in human-robot interactions inevitably becomes a key construct in the robot's formulation of subsequent actions. The role of both the robot and the user is to incrementally and continuously repair common ground (Clark et al., 1991; Clark and Krych, 2004). As such, physically situated robots need to be aware of the recurrently observed user states before planning future actions. Data-driven approaches are needed, that model the state of user uncertainty, confusion or hesitation. Robot tutors that provide instructions and guide humans in daily tasks, need to develop representations of their user's affective states and behavioural signals (D'mello et al., 2008; Kontogiorgos et al., 2020). If necessary, robot tutors should adapt their instruction strategies to the user levels of understanding (Kontogiorgos and Pelikan, 2020), and design their instructions for the recipient in the interaction (Pelikan and Broth, 2016).

Towards these efforts of modelling human behaviour for robot tutors, in this paper, we introduce a human-human guided task corpus. In the corpus, we are particularly interested in how humans collaborate in referential communication tasks. We asked subjects to instruct each other how to assemble IKEA furniture, without providing verbal instructions. Using the concept of an old children's game, known as *Chinese whispers*[1], subjects followed a chain of assembly instructions in uncontrolled dialogue, by taking the role of first the builder and then the instructor. We collected

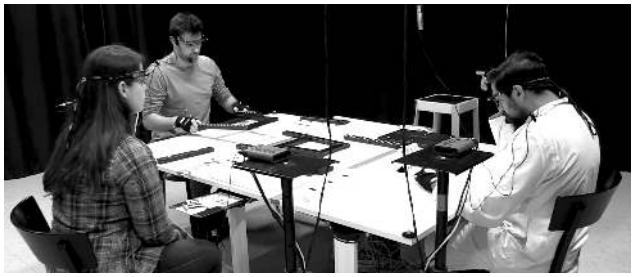

Figure 1: The collaborative assembly task. The instructor (on the right) guides the builder (on the left) how to assemble an IKEA stool.

multimodal data in speech, eye-gaze and pointing gestures, as well as actions in furniture piece movements, all automatically extracted using a multisensory setup (Jonell et al., 2018). These were processed and analysed in a first attempt to examine the effect of Chinese whispers. Do instructors influence subsequent instructors in their choice of verbal descriptors, in multimodal deictic signals (pointing, poising, etc.), or in paralinguistic components?

Examining the corpus further, we got interested in how speakers construct instructions in a collaborative nature, and how much influenced they are by their listeners' expressed signals of uncertainty or hesitations. We extracted high-dimensional features that represent the listeners' state of uncertainty and along with the speakers' verbal instructions, we were able to predict with high confidence that the speaker will repair a previous statement by reformulating their utterance (Kontogiorgos et al., 2019).

This corpus is of particular interest to researchers interested in how people, using the least-collaborative effort, establish, maintain and repair common ground (Clark et al., 1991; Clark and Wilkes-Gibbs, 1986). The dataset is only available for research purposes and not for commercial use. The data is anonymised, therefore no video or audio is available, but speech transcripts, eye-gaze and hand gesture labels, synchronised and represented in high dimensional

[1] also known as 'broken telephone' in some countries.

Figure 2: The interactions across all 34 sessions. The builder (interaction 12: on the right in yellow shirt) from each session becomes the instructor (interaction 13: on the left in yellow shirt) in the next session and a new builder is instructed how to do the task.

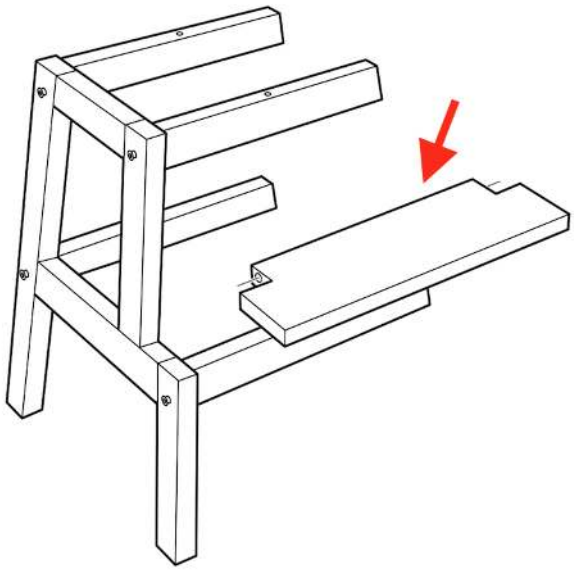

Figure 3: Instructors were given visual guides on furniture assembly without any verbal or textual instructions on what piece they should describe next. Image as shown in the IKEA 'Bekväm' stool catalogue.

features. Because of the fore-mentioned design decision, this corpus offers the possibility to model social robots' verbal and non-verbal behaviour in instruction and mediation tasks. To our knowledge, there is no other publicly available corpus which does the same.

2. Related work

The concept of Chinese whispers has been used in different domains, although infrequently. There are large opportunities in studying the effects of signal reconstruction after consequently introducing noise, particularly in collaborative dialogue. A message that is passing through different subjects, while developing a worse and worse signal-to-noise ratio, can ultimately develop no apparent connection to the originally constructed message (Wardy, 1993). While each channel is structurally or semantically close to its previous and next channels, noise is accumulated over time such that the modified message can potentially change beyond recognition from the original. In this corpus, noise is introduced by each new instructor, however within the limits of the guided assembly, allowing for consistency across different instructors.

Several multimodal corpora have been created over the last decade, using multisensory input, such as the ones described in (Carletta, 2007), (Mostefa et al., 2007), (Oertel et al., 2014), (Hung and Chittaranjan, 2010), (Oertel et al., 2013), (Stefanov and Beskow, 2016), (Kontogiorgos et al., 2018a). (Carletta, 2007) and (Mostefa et al., 2007) describe corpora collected in meetings, (Hung and Chittaranjan, 2010) and (Stefanov and Beskow, 2016; Kontogiorgos et al., 2018a) are examples of task-based scenarions such as games, and (Oertel et al., 2014) in job interviewing. (Oertel et al., 2013) in contrast to the corpora listed, gathers data from multi-party interactions 'in-the-wild', rather than constricted in-lab interactions.

An interesting corpus similar to the one presented here (Schreitter and Krenn, 2016) presents data with task-oriented dialogue and multimodal task descriptors, in different settings of guided tasks. Our design decision in introducing the chain effect was influenced by (Schreitter and Krenn, 2016). That corpus however, attempts to avoid the effect of Chinese whispers by occasionally introducing calibration sessions with an experimenter.

(Stefanov and Beskow, 2016) in their corpus, study the visual focus of attention of groups of participants, similarly to (Kontogiorgos et al., 2018a) in task-oriented dialogues. Both corpora include recorded groups of three participants while they engage in a task on a screen, but also in open-world dialogue (Bohus and Horvitz, 2009). Analyses on the contribution of different predictors and methods developed to predict listeners' visual focus of attention in multiparty interactions was developed based on these two corpora in (Stefanov et al., 2019).

3. Licence

4. Data Collection

4.1. Task

Subjects were instructed to assemble furniture. The task was to assemble a stool from IKEA as shown in Figure

3. Participants were sitting across each other and a table between them had all furniture pieces necessary for the assembly (Figure 4). The experiment setup also included additional pieces not used in the assembly, for distraction. To ensure variability in object descriptors, and further create instances of uncertainty, the setup included pieces similar to each other with variability in shape, size, colour, and location from the builder's perspective. To make each piece unique we created different patterns on the pieces using black and white tape. Each participant did the task twice, first as a builder, and then as an instructor (Figure 2).

In order to control for consistency in the referential communication task, but also leave space for 'noise' in the instructions, we used the concept of Chinese whispers. We started by teaching the first *instructor* how to do the task (researcher at KTH Royal Institute of Technology), and then asked him to introduce the task to the first *builder* participant. The builder then in the next session took the role of the instructor. By deciding to promote a chain effect in the assembly means we had little control over how the task would be performed. At the end of each task, the experimenters disassembled the IKEA stool and prepared for the next assembly. The assembly objects were therefore the same in every interaction and the builders were not aware of what they were about to build, or had previously seen the assembled stool. Additionally, the instructor was given a cheat sheet with a picture of the stool to indicate the order of their assembly.

During the assembly one of the experimenters was present, in the same room, facilitating the study but without participating in the assembly. The facilitator had knowledge of the assembly task and was present at all sessions. She pressed a button to play audio feedback on the task succession - signalling if the furniture was assembled as intended. The role was to ensure that the instructions are followed, to keep consistency in the assembly, but without interventions in the task or dialogue, leaving space for misunderstanding instructions and uncertainty, both natural consequences of the conversation. Subjects were informed that the facilitator would not interfere or help in any way in the assembly.

4.2. Dialogue

An illustration of uncertainty handling in instructions from a dialogue transcript (P20). The speaker continuously reformulates until they have established common ground.

INSTRUCTOR: *So the first one you should take*
BUILDER: *mhm*
INSTRUCTOR: *is the frame*
BUILDER: (Looks at table, moves hands)
INSTRUCTOR: *But the one with the stripes*
BUILDER: *okay*
BUILDER: (Looks towards left part of table)
INSTRUCTOR: *The black one*
BUILDER: (Looks at object)
INSTRUCTOR: *With the stripes*
BUILDER: (Reaches for object)
INSTRUCTOR: *Perfect*

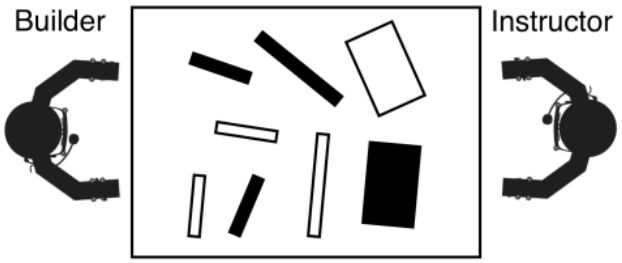

Figure 4: Subjects were sitting across each other to build the furniture and were wearing sensory equipment.

Due to the nature of the guided task, subjects are not able to establish grounding (Clark and Schaefer, 1989) until they have completed each step and carry on further in the task. All participants were instructing for the first time, which led to collaborative instructions, and naturally eliciting uncontrolled situations of uncertainty from the collaborators.

4.3. Participants

Data was collected from 34 participants. The facilitator was present in all 34 interactions, and was always the same experimenter. The mean age of the participants was 25.5 (SD 3.5) in range (21-39); 11 reported female and 23 male, and the majority of them were students or researchers at KTH Royal Institute of Technology in Stockholm, Sweden. All participants were fluent in English, with a mean 6.5 in scale 1-7 of self-reported English literacy. Participants reported little to no interference of sensory equipment in the task (eye-tracking glasses, microphone, gloves with motion capture markers) with 2.2 (in scale 1-7) in an equipment interference questionnaire item. All participants were experienced with digital technology (mean 6 in scale 1-7), and 20 our of 34 had interacted with a robot before (most common were Furhat, Yumi, Nao, Baxter, and Pepper). Participants also reported relatively experienced in assembling IKEA furniture (mean 4.6 in scale 1-7).

4.4. Procedure

Each session followed the same structure: The experimenter welcomed the builder and introduced them to the instructor, and thereafter introduced the experimental setup and helped them wear the sensory equipment. The instructor then (builder participant from previous session) started instructing the current builder how to assemble the furniture. At the end of the task, both builder and instructor filled up questionnaires on their experience from the task, in separate rooms. During the interaction we collected a fusion of multimodal data; a variety of input modalities that we combined to get information on participants' states and intentions. The very first instructor and last builder of the study was the same person, that helped starting and finishing the Chinese whisper chain. All participants signed a consent form for audio and video recording permissions and the ability to publish the results of the study. They were reimbursed with a cinema ticket.

Item	Builder	Instructor
Task coordination	How difficult did you find it to coordinate your behaviour with the instructor when you were working together?	How difficult did you find it to coordinate your behaviour with the builder when you were working together?
Task contribution	How much did the instructor contribute to the completeness of the task?	How much did the builder contribute to the completeness of the task?
Focus (self)	I remained focused on the instructor throughout our interaction.	I remained focused on the builder throughout our interaction.
Focus (other)	The instructor remained focused on me throughout our interaction.	The builder remained focused on me throughout our interaction.
Understanding	It was easy to understand the instructor.	The builder understood me.
Engagement (self)	How engaged were you during the task?	How engaged were you during the task?
Engagement (other)	How engaged was the instructor during the task?	How engaged was the builder during the task?
Task recall	How well do you remember the task if you were asked to do it again?	How well do you remember the task if you were asked to do it again?
Task difficulty	Was the task easy or difficult to build?	Was the task easy or difficult to instruct?

Table 1: Questionnaire items for builder and instructor (all in Likert 1-7 scale).

4.5. Corpus

The corpus consists of a total of 34 interactions. All recordings (mean length of assembly task: 3.8 minutes) contain data from various sensors capturing motion, eye gaze, gestures, and audio streams. Information on the aggregated and processed anonymised data from the corpus is available at the following webpage: `https://www.kth.se/profile/diko/page/material`.

4.6. Sensory data

Participants were wearing a pair of gloves with reflective markers and eye tracking glasses (Tobii[2] Glasses 2) which also had reflective markers on them. The room was surrounded with 17 motion capture cameras positioned in such a way that both gloves and glasses are always visible to the cameras. The participants were also wearing a close-talking microphone with input volume adjusted so that only their own voice is captured (Figure 4).

4.6.1. Motion capture

We used an OptiTrack motion capture system[3] to collect motion data from the subjects and the furniture pieces. The 17 motion capture cameras collected motion from reflective markers on 50 frames per second (manually adjusted to be in the same frame rate to the eye tracking glasses). To identify rigid objects in the 3d space we placed 4 markers per object of interest (glasses, gloves) and 6 per furniture piece, and captured position (x, y, z) and rotation (x, y, z, w) for each rigid object.

4.6.2. Eye gaze

In order to capture eye gaze in 3D space, we used the Tobii eye tracking glasses, so that we can accurately identify the gaze trajectory in space from the glasses' perspective. Combining the eye tracking data with the head motion (as captured by the motion capture system), we extracted eye-gaze data in 3d space using a real-time multisensory architecture (Jonell et al., 2018). Gaze samples were collected on 50 frames per second and the data was captured by tracking the subjects' pupil movements and a video from their point of reference. The facilitator was also wearing eye tracking glasses, however no gloves or microphone as she did not interfere in the assembly task or instructions.

4.6.3. Audio and video

Each participant's voice was recorded using channel separated close-talking microphones and transcribed in real-time using automatic speech recognition from the IBM Watson service [4]. There were two video cameras on a distance recording the interaction from different angles that we used for qualitative and conversational analyses.

4.6.4. Data processing

We used a real-time multisensory architecture (Jonell et al., 2018) to capture and sync all sensory data input and process to higher dimensional features represented in data such as proportional gaze to the conversational partner (builder or instructor) or gaze to any of the objects, and similarly, pointing to any of the objects available, as well as current words spoken, separated by keywords with timing information per word boundary.

The sensors we used required calibration in order to successfully capture motion and eye movements. We calibrated all 17 cameras positioning at the beginning of all recordings, while the eye tracking glasses required calibration on each recording per subject separately.

4.7. Subjective measures

At the end of each session, we gave both participants a questionnaire to measure their impression on the task and how collaborative were their efforts in the assembly (Table 1). Questionnaire items included: perceived coordination, task contribution, the focus of attention, mutual understanding, engagement, task recall, and task difficulty.

5. Analysis

5.1. Chinese whisper effect

The builder often reused the instructor's descriptive features, lexical choices, and deictic behaviour in their own subsequent instructions. Out of 55 descriptive words, 20 occurred in more than four sessions. Chi-square tests revealed that 6 of these words were repeated in at least three consecutive sessions, e.g. 'dots' and 'stick': ['dots':$\chi^2 = 8.84, p = .002$], ['stick':$\chi^2 = 13.84, p < .001$]. Similarly, terms that described shapes of objects occurred in at least three consecutive sessions e.g. 'stripes' and 'circles': ['stripes':$\chi^2 = 6.05 p = .013$], ['circles':$\chi^2 = 11.09, p < .001$].

Many words were reused in long chains across sessions before being replaced by others. We found the same chain effect for pointing during the verbal descriptions. By using a chain of builders turning into instructors across 34 sessions, we could analyse the 'stickiness' of descriptive words. For example, words like 'rod', 'thing' and 'plank' were never

[2] `http://www.tobiipro.com/`

[3] `http://optitrack.com/`

[4] `https://www.ibm.com/watson`

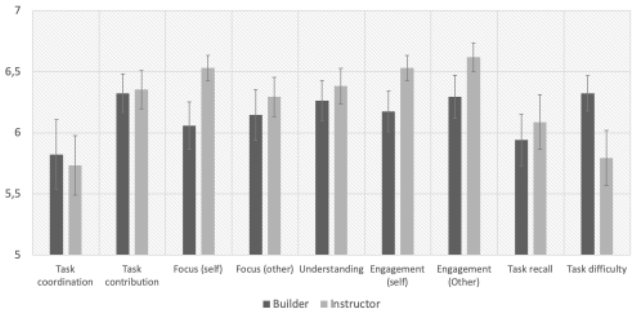

Figure 5: Self-reported measures in task experience from the builder and instructors' point of view. Error bars indicate standard error of the mean (n=34).

Item	Builder Mean	Builder SD	Instructor Mean	Instructor SD
Task coordination	5.8	1.6	5.7	1.4
Task contribution	6.3	0.9	6.3	0.9
Focus (self)	6.0	1.1	6.5	0.6
Focus (other)	6.1	1.2	6.2	0.9
Understanding	6.2	0.9	6.3	0.8
Engagement (self)	6.1	0.9	6.5	0.6
Engagement (other)	6.2	1.0	6.6	0.6
Task recall	5.9	1.2	6.0	1.3
Task difficulty	6.3	0.8	5.7	1.3

Table 2: Questionnaire data on participants' experience from the task. Mean and standard deviations (N=34).

picked up by builders, while referring language strategies such as indicating the object location, or using certain lexical forms were often passed on. Some words were often reintroduced, indicating that they might be the most obvious way to describe the assembly objects. Thus, the chain effect seems to be an appropriate way to get a composition of lexical variations and conceptual pacts among subjects and across assembly sessions.

5.2. Subjective measures

Using the self-reported questionnaire items, we report the means and standard deviations for each of the items in perceived coordination, task contribution, focus of attention, mutual understanding, engagement, task recall, and task difficulty in Table 2 and Figure 5. The measures are grouped together to indicate differences between builders and instructors.

One-way ANOVAs on the role of the participant (builder or instructor) showed a significant effect on the focus (self) measure: $F(1,67) = 4.571$; $p = .036$, and a marginally significant effect on the task difficulty measure: $F(1,67) = 3.882$; $p = .053$. Builders thought they were less focused on the instructor (as they were focused on the task), than what the instructors thought about themselves being focused on the builder. But builders also thought the task was more difficult than what the instructors thought it was. No other significant effects were found.

5.3. Effect of modalities on reference resolution

In previous work (Kontogiorgos et al., 2018b), we used the current corpus to analyse which modalities, or a combination thereof, carry the most informative cues for resolving referring expressions in the collaborative task. The study was designed with two goals: (i) identify the most salient object of attention during each step of the assembly task, and (ii) bring insight in the reliability of each of the verbal and non-verbal cues that a human observer uses to identify errors. The saliency of an object was defined as the proportion of gaze fixations or pointing gestures during a referring expression. In other words, the longer the object is looked at or pointed at, the more probable it is to be the target object of the current interaction segment.

We trained a binary-choice SVM classifier with radial-basis-function kernel that combined the instructor's speech with gaze, head, and pointing gestures of both the instructor and the builder for identifying the saliency of each object. On a 5-fold cross-validation, we evaluated the performance of the classifier by calculating the mean ranking of objects' saliencies from the probabilistic output of the classifier (the higher the confidence of the classification, the higher the rank). Out of all combinations, the most effective results were shown on the classifier trained only on the instructor's speech and head movements at 88%, closely followed by instructor's speech with gaze fixations at 85%, while the unimodal classifiers performed at the 58-66% accuracy rate. Such results showed that while a multimodal approach significantly outperforms unimodal, the noisiness of non-verbal cues, if not handled properly, can decrease the effectiveness of the prediction.

5.4. Estimating listener uncertainty

In another application of this corpus we investigated listener signals to estimate uncertainty (Kontogiorgos et al., 2019). Using automatically extracted non-verbal cues from the builder's gaze and pointing gestures, we aimed to predict if the instructor will reformulate their utterances. Instructor repairs and reformulations were time-segmented manually to indicate intrinsic builder signals such as the ones of uncertainty and hesitation. Our assumption was that in the continuous effort to establish grounding, speakers reformulate messages, if necessary, as shown in listeners signals of uncertainty. Using the manually time-segmented instruction units, we asked human annotators to indicate if the listener looked uncertain during the speaker instructions.

We also trained a Random Forest classifier to classify the same instruction units into two classes of uncertainty and non-uncertainty. Our results showed that a RF classifier, based on non-verbal cues, outperformed human annotators with mean accuracy 79% and 72% respectively. These findings indicated that using listener signals is a fundamental construct in speakers' decisions to reformulate their utterances. During the building task, conversational partners establish common ground using pragmatic feedback; both the speaker and the listener reformulate their references until they feel they are understood.

6. Discussion

In this paper, we presented a multimodal dataset collected in a collaborative task, where one participant instructed another how to assemble an IKEA stool. The dataset contains speech, eye-gaze and pointing gestures along with object movements. The corpus is of particular interest to researchers who focus on behavioural analyses of interaction. It is also interesting to researchers that build data-driven

models of multimodal human behaviour in situated dialogue, particularly in applications of language grounding, disambiguation of referring expressions and intent recognition for human-robot interactions.

The 'Chinese Whispers' approach that we employed, in order to avoid introducing implicit experimenter biases into participants' behaviour, can be of interest for further examination. More specifically, how does the carried out effect propagate across the pair of participants, which behavioural patterns are more likely to be replicated, and how does this ripple effect disappear or reinforce itself with time?

We also found differences in how builders and instructors perceived allocating their focus towards each other. Builders seemed to be focusing more on the task rather than the instructor and also thought their task was difficult in comparison to the instructors. The instructors however, thought they focused more on the builders, as they have to supervise their assembly efforts and ensure builders are following correctly their instructions. They also thought the task was easier when instructing, rather than when assembling. This is not surprising, as in task-oriented dialogues, conversational partners' focus changes throughout the interaction between each other and the task. We also found that both builder and instructor, thought they contributed equally to task completeness, showing that the roles of performing the task and guiding the task are equally important.

While we made an attempt to not introduce biases in verbal descriptions of furniture objects, it may be possible that the stimuli presented to subjects heavily affect their word selections. There was little variability in objects, outside of the context of the assembly task. As such, it is likely that any chain effects found may not generalise across different domains of tasks, or where referential communication is not relevant.

Another limitation of the presented corpus is the result of the restricted, task-oriented dialogue of collaborative assembly. Even though participants were free to use any verbal or non-verbal cues to interact, the assembly task with predefined sequence of steps and distribution of rigid participant roles, may have constricted the interaction in such a way, that it will be unclear whether models developed based on this dataset may be generalised to open-ended conversational types of interactions. Nevertheless, the intrusiveness of the sensors that we used for data collection, namely eye tracking glasses and motion tracking gloves, might also have influenced participants' non-verbal behaviour, even if they attribute no interference to the task. Similarly, the presence of an experimenter in the room may have affected subjects' behaviours as well.

7. Conclusion

On a final note, we believe the presented dataset, establishes a valuable contribution to the research community, and we have shown how the dataset can be used both in cases when subjects establish common ground, but also in cases when common ground is not satisfied, but continuously awaited. The presented dataset has implications to the design of artificial agents expected to guide humans or teach them how to do certain tasks. We encourage other researchers to further explore this experimental design, and nevertheless the effects of Chinese Whispers in human communication and human-machine interactions.

Acknowledgements

This work was supported by the Swedish Foundation for Strategic Research project FACT (GMT14-0082). We would like to thank all the participants that took part in the study, and Per Fallgren for taking part in data collection. We would also like to thank Simon Alexandersson for helping in processing motion capture and eye-tracking data. Finally, we would like to thank Hannah Pelikan for transcribing and analysing parts of the corpus.

References

Bohus, D. and Horvitz, E. (2009). Open-world dialog: Challenges, directions, and prototype. In *Proceedings of IJCAI'2009 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Clark, H. H. and Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of memory and language*, 50(1):62–81.

Clark, H. H. and Schaefer, E. F. (1989). Contributing to discourse. *Cognitive science*, 13(2):259–294.

Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1):1–39.

Clark, H. H., Brennan, S. E., et al. (1991). Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149.

D'mello, S. K., Craig, S. D., Witherspoon, A., Mcdaniel, B., and Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User modeling and user-adapted interaction*, 18(1-2):45–80.

Hung, H. and Chittaranjan, G. (2010). The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 879–882. ACM.

Jonell, P., Bystedt, M., Fallgren, P., Kontogiorgos, D., Lopes, J., Malisz, Z., Mascarenhas, S., Oertel, C., Raveh, E., and Shore, T. (2018). FARMI: A Framework for Recording Multimodal Interactions. In *Language Resources and Evaluation Conference LREC 2018*.

Kontogiorgos, D. and Pelikan, H. (2020). Towards adaptive and least-collaborative-effort social robots. In *International Conference on Human Robot Interaction (HRI)*.

Kontogiorgos, D., Avramova, V., Alexandersson, S., Jonell, P., Oertel, C., Beskow, J., Skantze, G., and Gustafsson, J. (2018a). A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction. In *LREC*.

Kontogiorgos, D., Sibirtseva, E., Pereira, A., Skantze, G., and Gustafson, J. (2018b). Multimodal reference resolution in collaborative assembly tasks. In *Proceedings of the 4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, pages 38–42. ACM.

Kontogiorgos, D., Pereira, A., and Gustafson, J. (2019). Estimating uncertainty in task-oriented dialogue. In *2019 International Conference on Multimodal Interaction*, pages 414–418. ACM.

Kontogiorgos, D., Abelho Pereira, A. T., Sahindal, B., van Waveren, S., and Gustafson, J. (2020). Behavioural responses to robot conversational failures. In *International Conference on Human Robot Interaction (HRI)*.

Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S. M., Tyagi, A., Casas, J. R., Turmo, J., Cristoforetti, L., Tobia, F., et al. (2007). The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language Resources and Evaluation*, 41(3-4):389–407.

Oertel, C., Cummins, F., Edlund, J., Wagner, P., and Campbell, N. (2013). D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7(1-2):19–28.

Oertel, C., Funes Mora, K. A., Sheikhi, S., Odobez, J.-M., and Gustafson, J. (2014). Who will get the grant?: A multimodal corpus for the analysis of conversational behaviours in group interviews. In *Proceedings of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, pages 27–32. ACM.

Pelikan, H. R. and Broth, M. (2016). Why that nao? how humans adapt to a conventional humanoid robot in taking turns-at-talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4921–4932.

Schreitter, S. and Krenn, B. (2016). The ofai multi-modal task description corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1408–1414.

Stefanov, K. and Beskow, J. (2016). A multi-party multimodal dataset for focus of visual attention in human-human and human-robot interaction. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016, 23-28 of May*. ELRA.

Stefanov, K., Salvi, G., Kontogiorgos, D., Kjellström, H., and Beskow, J. (2019). Modeling of human visual attention in multiparty open-world dialogues. *ACM Transactions on Human-Robot Interaction (THRI)*, 8(2):8.

Thomaz, A., Hoffman, G., Cakmak, M., et al. (2016). Computational human-robot interaction. *Foundations and Trends® in Robotics*, 4(2-3):105–223.

Wardy, R. (1993). Chinese whispers. *The Cambridge Classical Journal*, 38:149–170.