A Multimodal Corpus for Mutual Gaze and Joint Attention in Multiparty Situated Interaction

Dimosthenis Kontogiorgos[1], Vanya Avramova[1], Simon Alexanderson[1], Patrik Jonell[1], Catharine Oertel[1 2], Jonas Beskow[1], Gabriel Skantze[1], Joakim Gustafson[1]

[1] KTH Royal Institute of Technology, Sweden
[2] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Abstract

In this paper we present a corpus of multiparty situated interaction where participants collaborated on moving virtual objects on a large touch screen. A moderator facilitated the discussion and directed the interaction. The corpus contains recordings of a variety of multimodal data, in that we captured speech, eye gaze and gesture data using a multisensory setup (wearable eye trackers, motion capture and audio/video). Furthermore, in the description of the multimodal corpus, we investigate four different types of social gaze: referential gaze, joint attention, mutual gaze and gaze aversion by both perspectives of a speaker and a listener. We annotated the groups' object references during object manipulation tasks and analysed the group's proportional referential eye-gaze with regards to the referent object. When investigating the distributions of gaze during and before referring expressions we could corroborate the differences in time between speakers' and listeners' eye gaze found in earlier studies. This corpus is of particular interest to researchers who are interested in social eye-gaze patterns in turn-taking and referring language in situated multi-party interaction.

Keywords: multimodal situated interaction, social eye-gaze, referential gaze, joint attention, mutual gaze, reference resolution

1. Introduction

In this corpus we combine verbal and non-verbal communication cues to model shared attention in situated interaction. We capture multimodal cues using a multisensory setup in order to extract information on visual attention during deictic references in collaborative dialogue. We made use of a moderator who had the task of leading the interaction and making sure that both participants were involved in the task and in the conversation. During the interaction the moderator used referring expressions and gaze to direct the participants' attention to objects of interest on a large touch screen (figure 1).

Figure 1: Participants used referring expressions and gaze to direct each other's attention while moving objects on a large display.

Our purpose of collecting the corpus presented in this paper is to study social eye gaze in multiparty interaction. The goal is to create visual attention models that make it possible for our robots to direct humans' attention to certain objects. There are several novelties with our corpus: Firstly, it is a three-party interaction with and without an interactive touch screen where we have synchronised data streams of gaze targets, head direction, hand movement and speech (run through ASR with word timings). Second, one of the three participants is a mediator that has the role of encouraging the participants to reconsider their decisions and to foster collaboration. In all recordings we use the same person as the mediator, which makes it possible for us to build coherent verbal and non-verbal behavioural models. We will use these to develop a robot that can be used as a moderator in similar collaborative tasks. To our knowledge there is no other publicly available corpus with these features.

During their interaction, the participants collaborated to furnish a virtual apartment using a collection of available furniture objects. It was their task to discuss and decide on which objects they would use, given that they had a limited budget. The moderator had the role of leading the discussion. However, this does not change the fact that this is an example of dynamic multiparty situated interactions that (Bohus and Horvitz, 2009) define as an open-world dialogue. When discussing the furniture objects, the participants naturally made use of a combination of verbal referring expressions and non-verbal cues such as deictic gestures and referential gaze. We are particularly interested in the participants' gaze behaviour just before and during verbal referring expressions. When analysing our corpus we find that listeners and speakers display different visual behaviour in some cases.

In the next sessions, we provide an overview of the state-of-the-art in multimodal multiparty corpora and relevant works in various types social eye gaze. We further describe our process in data collection and experiment design, and our methods for automatic eye gaze extraction. Finally, we go through the collected data and give an overview of the participants' gaze behaviour during the task-oriented dialogues and our findings supported by the relevant literature.

2. Background

2.1. Multiparty multimodal corpora

Over the last decade, more and more multimodal multiparty corpora have been created, 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). (Carletta, 2007) and (Mostefa et al., 2007) fall into the category of meeting corpora, (Hung and Chittaranjan, 2010) and (Stefanov and Beskow, 2016) are examples of games corpora, and (Oertel et al., 2014) a job interviewing corpus. The corpus from (Oertel et al., 2013) in contrast to the corpora listed above tries to escape the lab environment and gathers data of multi-party interactions "in the wild".

In (Carletta, 2007), (Mostefa et al., 2007) and (Oertel et al., 2013) the visual focus of interlocutors is divided; i.e. there are stretches in the corpus in which interlocutors' main focus of attention is on each other and there are stretches in the corpus where they mainly focus on e.g. the white board or sheets of papers which are laying in front of them. However, with the technical set-up used at the recordings at the time, it was hard to infer the exact points the participant were looking. In (Oertel et al., 2014) and (Hung and Chittaranjan, 2010) participants' visual focus of attention is solely focused on the other participants (no other objects are present during the recordings).

Another example is (Stefanov and Beskow, 2016) who, in their corpus, study the visual focus of attention of groups of participants. For this, they recorded groups of three participants while they were sorting cards. They also recorded groups of three participants, discussing their travel experiences without any objects present that might distract the visual focus of attention.

Finally, while (Lücking et al., 2010) is not a multiparty corpus, it should be mentioned here as it is similar to the corpus described in this paper, particularly well suited for the study of referring expressions. In terms of experimental setup, the corpus recording described in this paper is most similar to (Stefanov and Beskow, 2016).

2.2. Social eye-gaze

Social eye gaze refers to the communicative cues of eye contact between humans and is usually referred to by 4 main types (Admoni and Scassellati, 2017): 1. Mutual gaze where both interlocutors' attention is directed at each other, 2) Joint attention where both interlocutors focus their attention on the same object or location, 3) Referential gaze which is directed to an object or location and often comes together with referring language and 4) Gaze aversions that typically avert from the main direction of gaze - i.e. the interlocutor's face.

Joint attention is of particular importance for communication. It provides participants with the possibility to interpret and predict each other's actions and react accordingly. In the current corpus for example, the modeling of joint attention is of particular importance as it provides participants with the possibility to track the interlocutors' current focus of attention in the discourse (Grosz and Sidner, 1986). A common quality of joint attention is that it may start with mutual gaze to establish where is the attention of the interlocutor and end towards the referential gaze direction to the most salient object (Admoni and Scassellati, 2017).

Also, as participants are very likely to be focused more on the display than each other, joint attention will be crucial to discern whether they are paying attention to each other. Modelling of joint attention is also crucial when developing fine-grained models of dialogue processing (Schlangen and Skantze, 2009), which for example makes it possible for a dialogue system to give more timely feedback (Meena et al., 2014). With regards to multi-party interaction there are also recent studies which model the situation in which the interaction takes place, in order to manage several users talking to the system at the same time (Bohus and Horvitz, 2010), and references to objects in the shared visual scene (Kennington et al., 2013).

2.3. Multimodal reference resolution

A reference is typically a symbolic representation of a linguistic expression to a specific object or abstraction. During early attempts in verbal communications between humans and machines researchers used rule-based systems to disambiguate words and map them to referent objects in virtual worlds (Winograd, 1972). Research has also focused in disambiguating language using multimodal cues; starting in the late 70s with Richard Bolt's "Put-That-There" (Bolt, 1980), to recent approaches using eye-gaze (Mehlmann et al., 2014; Prasov and Chai, 2008; Prasov and Chai, 2010), head pose (Skantze et al., 2015), and pointing gestures (Lücking et al., 2015). Gross et. al recently explored the variability and interplay of different multimodal cues in reference resolution (Gross et al., 2017).

Eye gaze and head direction have been shown to be good indicators of object saliency in human interactions; researchers have developed computational methods to construct saliency maps and identify humans' visual attention (Bruce and Tsotsos, 2009; Sziklai, 1956; Borji and Itti, 2013; Sheikhi and Odobez, 2012). Typically speakers direct the listeners' attention to objects using verbal and non verbal cues and listeners often read the speaker's visual attention during referring expressions to get indications on the referent objects.

3. Data collection

3.1. Scenario

We optimised for variation in conversational dynamics by dividing the current corpus recordings into two conditions. In the first condition the moderator facilitated a discussion about participants' experience on the topic of sharing an apartment. In this condition no distracting objects were existent and the screen on the large display was off. In the second condition the participants were asked to collaborate on decorating an apartment (Figure 2) that they should imagine they would be moving in together. They were given an empty flat in which they had to decide where to place furniture which they could buy from the store. The stores were provided as extra screens on the application, that they could go through to get new pieces of furniture. They were given a limited budget which fostered their decisions and discussions on what objects to choose. Given the variety of available objects they would need to compromise their choices and collectively decide where to place objects.

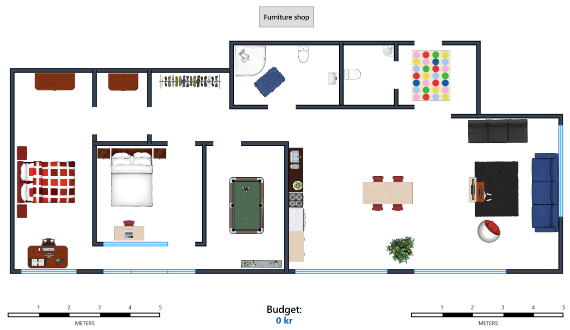

Figure 2: The application running on the large display. The objects were ”bought” from the furniture shop that had a variety of available furniture.

3.2. Participants

We collected data from 30 participants with a total of 15 interactions. Our moderator (native US-English speaker) was present on all 15 sessions facilitating the structure of the interactions and instructing participants on their role for completing the task. The moderator was always at the same part of the table and the two participants were sitting across the moderator (Figure 1). The mean age of our participants was 25.7; 11 were female and 19 were male and the majority of them were students or researchers at KTH Royal Institute of Technology in Stockholm, Sweden.

3.3. Corpus

Our corpus consists of a total of 15 hours of recordings of triadic interactions. All recordings (roughly one hour each) contain data from various sensors capturing motion, eye gaze, audio and video streams. Each session follows the same structure: The moderator welcomes the participants with a brief discussion on the topic of moving in a flat with someone, and thereafter introducing the setup and scenario of the two participants planning their moving in the same apartment using the large display. During the interactions we collected a set of multimodal data: a variety of input modalities that we combined to get information on participants' decision making and intentions.

Out of the 15 sessions, 2 sessions had no successful eye gaze calibration and were discarded. One session had synchronisation issues on the screen application which was also discarded. Last, one session had large gaze data gaps and was discarded as well. The rest 11 sessions of the aggregated and processed data from the corpus are further described on the rest of the paper and are available at: `https://www.kth.se/profile/diko/page/material`.

3.4. Experimental setup

Participants were situated around a table which had a large touch display. On the display there was an application we developed to facilitate their planning of a moving in together in a flat scenario. We gave participants a pair of gloves with reflective markers and eye tracking glasses (Tobii Glasses 2[1]) which also had reflective markers on to track their position in space. The room was surrounded with 17 motion capture cameras positioned in such a way that both gloves and glasses are always on the cameras' sight.

The moderator was also wearing eye tracking glasses. Since our aim is to develop models for a robot without hands, the moderator was not wearing gloves with markers and was instructed to avoid using hand gestures. There were two cameras placed on the table capturing facial expressions and a regular video camera at a distance recording the full interaction for annotation purposes. On the glasses we also placed lavalier microphones (one per participant), in such a way that we capture the subject's voice separated from the rest of the subjects' speech and with a volume consistency.

The participants collaborated in the given scenario for 1 hour where they discussed and negotiated to form a common solution in apartment planning. At the end of the recording, we asked participants to fill a questionnaire on their perception of how the discussion went, their negotiations with the other participant and finally some demographic information. All participants were reimbursed with a cinema ticket.

3.5. Motion capture

We used an OptiTrack motion capture system[2] to collect motion data from the subjects. The 17 motion capture cameras collected motion from reflective markers on 120 frames per second. To identify rigid objects in the 3d space we placed 4 markers per object of interest (glasses, gloves, display) and captured position (x, y, z) and rotation (x, y, z, w) for each rigid object. While 3 markers are sufficient for capturing the position of a single rigid body, we placed a 4th marker on each object for robustness. That way, if one of the markers was not captured we would still identify the rigid object in space.

3.6. Eye gaze

We were interested in collecting eye gaze data for each participant in order to model referential gaze, joint attention, mutual gaze and gaze aversion in multiparty situated dialogue (figure 3). In order to capture eye gaze in 3D space we used eye tracking glasses with motion capture markers. This made it possible to accurately identify when a participant's gaze trajectory intersected objects or the other interlocutors. It also made it possilbe to capture gaze aversion, i.e. when participants gazed away when speaking or listening to one of the participants.

Gaze samples were on 50Hz and the data was captured by tracking the subjects' pupil movements and pupil size and a video from their point of reference. We placed reflective markers on each pair of glasses, and were therefore able to identify their gaze trajectory in x and y on their point of reference and then resolve it to world coordinates using the glasses' relevant position in 3d space.

The glasses using triangulation from both eyes' positions, also provided a z value, which would refer to the point

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

where the trajectories from the two eyes meet. That point in space (x, y, z) we used to resolve the eye gaze trajectory in 3d space.

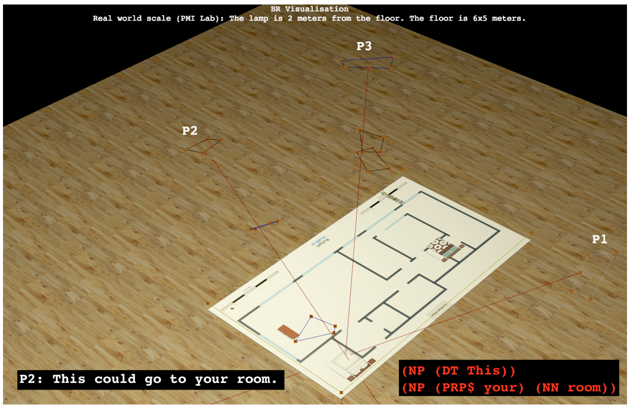

Figure 3: By combining synchronised data from motion capture and eye tracking glasses, we captured the participants' eye gaze in 3d space but also head pose and gestures.

3.7. Speech and language

We recorded audio from each participant using channel separated lavalier microphones attached to the eye tracking glasses. Each microphone captured speech that we later used to automatically transcribe using speech recognition and resolve the spoken utterances to text. We used a voice activity detection filter in all channels to separate captured speech from other speakers. We further used Tensorflow's[3] neural network parser to form syntactic representations of the automatically transcribed natural language text.

3.8. Facial expressions

In order to extract facial expressions we placed GoPro cameras in front of the participants and moderator on the table. We also recorded each session from a camera placed on a tripod further in the room to capture the interaction as a whole and for later usage in annotating the corpus.

3.9. Application

We built an application to enable the interaction with several virtual objects (figure 2). The application consisted of two main views: a floor plan and a furniture shop. The floor plan displayed on the large multi-touch display was used by the participants during their interactions as the main area to manipulate the objects while the shops were used to collect new objects.

The shop screens were divided by room categories such as kitchen or living room. They naturally induced deictic expressions as the participants referred to the objects during their negotiations using both language and referential gaze (i.e. "it", the desk", "the bed"). The floor plan on the contrary was the main display were they could manipulate objects by placing them in rooms and used referring language to these virtual locations (i.e. "here", "there", "my room").

[3]https://www.tensorow.org/versions/r0.12/tutorials/syntaxnet/

The floor plan view also displayed the apartment scale, and the budget the participants could spend for objects. The budget was limited and, after initial pilots, it was decided to have a value that would be enough to satisfy both participants' furniture selections but also limited enough to foster negotiations on what objects they would select. They were also allowed to only choose one of each item in order to foster negotiations further.

The application maintained event-based logging at 5 fps to keep track of the interaction flow, object movement and allocations. Each event carried time stamps and relevant application state, as well as positions and rotation data for all objects. By placing markers in the corners of the screen it could be placed into the motion capture coordinate system. This allowed us to get the virtual object events in 3D space and capture the participants' visual attention to these objects.

3.10. Subjective measures

At the end of each session we gave participants a questionnaire to measure their impression on how the discussion went and how well they thought they had collaborated when decorating the appartment. We used these measures to measure dominance, as well as collaborative behaviour and personality. All participants were asked to fill personality tests before coming to the lab; the tests included introversion and extroversion measures (Goldberg, 1992).

3.11. Calibration

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 subject separately. That is due to the fact that each participant's eye positions vary but also how their eyes move during saccades.

4. Annotations

We annotated referring expressions to objects on the display by looking at what object the speaker intended to refer to. Speakers typically drew their interlocutors' attention to objects using deictic expressions; referring language, referential gaze, as well as mutual gaze. For every dialogue there was a set of references to objects not included in our simulated environment -"I also have a fireplace in my flat, but I do not use it a lot". However, we only annotated references that can be resolved to objects existing in our simulated environment, i.e. the application on the touch display. The references we can resolve in this environment are therefore only a subset of all possible references in the dialogue.

For the annotations we used the videos of the interactions together with the ASR transcriptions and the current state of the app (visible objects or current screen on the large display). We defined each referring expression by looking at the time of the speaker's utterance. The timing was defined as from the ASR transcriptions that were synchronised with the gaze and gesture data. Utterances were split into inter pausal units (IPUs) that were separated by silent pauses longer than 250ms (Georgeton and Meunier, 2015).

The ASR transcriptions were used as input to the Tensorflow syntactic parser where we extracted the part-of-speech (POS) tags. Similarly to (Gross et al., 2017), as linguistic indicators we used full or elliptic noun-phrases such as ”the sofa”, ”sofa”, ”bedroom”; pronouns such as ”it” or ”this” and ”that”; possessive pronouns such as ”my”, ”mine”; and spatial indexicals such as ”here” or ”there”. The ID of the salient object or location of reference was identified and saved on the annotations. In some cases there was only one object referred to by the speaker, but there were also cases where the speaker referred to more than one object in one utterance (i.e. ”the table and the chairs”). The roles of the speaker and listeners varied in each expression. In total we annotated 766 referring expressions out of 5 sessions, roughly 5 hours of recordings which was about one/third of our corpus.

5. Data processing

The variety of modalities was post-processed and analysed, where in particular combined eye gaze and motion capture provided an accurate estimation of gaze pointing in 3D space. We used a single-world coordinate system in meters on both eye gaze and motion capture input streams and we identified the gaze trajectories, head pose and hand gestures. We started with a low-level multimodal fusion of the input sensor streams and further aggregated the data together with speech and language to high-level sequences of each participant's speech and visual attention. Each data point had automatic speech transcriptions, syntactic parsing of these transcriptions, gesture data in the form of moving objects and eye gaze in the form of target object or person.

5.1. Data synchronisation

We used sound pulses from each device to sync all signals in a session. The motion capture system sent a pulse on each frame captured, while the eye tracking glasses sent 3 pulses every 10 seconds. The application also sent a pulse every 10 seconds. All sound signals were collected on the same 12 channel sound card which would then mark the reference point in time for each session.
Audio signals were also captured on the soundcard, therefore we were able to identify a reference point that would set the start time for each recording. That was the last of the sensors to start, which was one of the eye tracking glasses. Apart from the sound pulses, we used a clapperboard with reflective markers on its closed position which would mark the start and the end of each session. We used that to sync the video signals and as a safety point in case one of the sound pulses failed to start. Since the application on the display sent sync signals, we used it to mark the separation in time of the two experimental conditions. The first one (free discussion) ended when the app started which would lead to the second condition (task-oriented dialogue).

5.2. Visualisation and visual angle threshold

As illustrated in figure 3 we calculated the eye gaze trajectories by combining motion capture data and eye tracking data per frame. We implemented a visualisation tool[4] in WebGL to verify and evaluate the calculated eye gaze in 3d space. Using this tool we were able to qualitatively investigate the sections of multimodal turn taking behaviour of participants by visualising their position and gaze at objects or persons, along with their transcribed speech and audio in wav format. We also used this tool to empirically define the visual angle threshold for each eye tracker on the accuracy of gaze targets. During pilots we asked participants to look at different objects ensuring there are angle variations to identify the threshold of the gaze angle to the dislocation vector (between the glasses and the salient object).

5.3. Automatic annotation of gaze data

We extracted eye-gaze data for all participants and all sessions and noticed that there were gaps between gaze data points quite often and in some sessions more than in others. We applied smoothing to eliminate outliers and interpolated gaps in a 12 frame step (100ms on our 120 fps data). Typically an eye fixation is 250ms but no smaller than 100ms (Rayner, 1995) which defined our smoothing strategy. The data was then filtered in the same time window to only provide data points with fixations and not include saccades or eye blinks.
There were however, gaps that were longer than 12 frames, caused by lack of data from the eye trackers. In such cases the eye trackers had no information on the eyes' positions which means that there was no knowledge on if a speaker/listener looked at the referent object. In such cases of no gaze data, we went through the manually annotated referring expressions and checked the relevant frames of the gaze data. If at least one of the participants had no gaze data, we would discard the relevant referring expression from our analysis. After cleaning those cases we had remaining 582 expressions out of the 766 initially annotated[5] The average error of gaze data loss we had was 40.8%. The session with the max gaze error rate was 71.5% while the min was 26.9%.
During an referring expression participants' visual attention spanned through a) a variety of visible objects on the screen, b) their interlocutors or c) none of the above which we assumed on this corpus to be gaze aversion. The prominent objects of visual attention were identified by calculating each participant's visual angle α, between their gaze vector g to the dislocation vector o for every visible object on the screen for every frame.

$$\alpha_{ij} = \arccos\left(\frac{\vec{g_{ij}} \cdot \vec{o_{ij}}}{|\vec{g_{ij}}|\,|\vec{o_{ij}}|}\right) \tag{1}$$

Similarly, we approximated each person's head with a sphere with radius of 0.2m, and automatically annotated all frames where the gaze vector intersected the sphere as visual attention towards that person.
Finally, after filtering for eye fixations, we calculated the proportional gaze likelihood per object:

$$P(o_i \mid t_i) = \frac{c(o_i, t_i)}{c(t_i)} \tag{2}$$

[4]Available at https://www.kth.se/profile/diko/page/material

[5]The 5 sessions chosen for annotation were the ones with the smallest percentage of gaze data loss.

For each object o_i we gathered the proportional gaze data points and counted the amount of time the object was gazed during the time t of an utterance. As a gaze target we defined the area around any object that is within the threshold of the visual angle defined above. In many cases more than one objects competed for saliency in the same gaze target area. We therefore defined as joint attention not the group of objects gazed by the interlocutors but the gaze at the same area around the prominent object given the visual angle threshold.

6. Results

In the current section, we describe gaze patterns as observed between and across the two conditions, and proportional gaze during and before referring expressions. In figure 4, we show the proportional eye-gaze of the moderator as well as the participants' eye gaze on the display. As can be observed, the participants spent more time gazing on the display than at the moderator or at each other, which was expected to observe during a task-oriented dialogue.

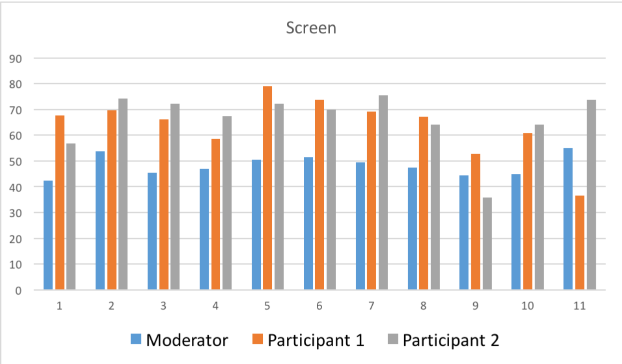

Figure 4: Proportional amount of participant eye-gaze per session on the display during the second condition (task-oriented dialogue).

For each utterance a set of prominent objects was defined from the annotations. Given the gaze targets per participant, we calculated the proportional gaze from the speaker and the listeners during the time of the utterance and exactly 1 second before the utterance. Since all interactions were triadic, there were two listeners and one speaker at all times. To compare across all utterances, we looked at the mean proportional gaze of the two listeners to the area of the prominent object to define them as the listener group. We then compared the gaze of the listener group to the speakers gaze close to the referent objects. We also looked at the proportional combined gaze to other objects during the utterance (all other objects that have been gazed at during the utterance), gaze at the speaker and averted gaze. In figure 5, the blue colour refers to the proximity area of the referent object from the speaker's utterance while orange refers to the gaze at all other known objects combined on the virtual environment. Grey is for the gaze to the listeners or the speaker during the utterance and yellow for averted gaze.

6.1. Eye gaze during referring expressions

We looked at all references to objects (N = 582) and compared the means of the proportional gaze of the speaker to the proportional listeners' gaze to the proximity area of the referent objects, the rest of the gazed objects and gaze towards each other. We conducted paired sample t-tests and found significant difference between the speaker gaze to the proximity area to the referent object (M = 46.69, Std.Error = 1.37) against the listener gaze on the area around the referent object (M = 39.45, Std.Error = 1.11) with [t = 4.942, p = 0.001]. There was no significant difference however on the speaker's gaze to the listeners against the gaze from the listeners to the speaker [t = -0.816, p = 0.415] (mutual gaze).

There was also a significant difference on the proportional gaze of the speaker to the area around the referent object (M = 46.69, Std.Error = 1.37) against the proportional gaze to other objects during the same utterances (M = 32.41, Std.Error = 1.25), [t = 5.887, p = 0.001]. However, no significant difference was found on the same case for the listeners' gaze [t = -0.767, p = 0.444].

6.2. Eye gaze before referring expressions

Previous studies have revealed that speakers typically look at referent objects about 800-1000ms before mentioning them (Staudte and Crocker, 2011). We therefore looked at 1 second before the utterance for all utterances (N = 582) and the gaze proportions around the area to the object(s) that were about the be referred to. We compared the speaker's gaze to the referent objects to the listeners' gaze using paired sample t-tests. The speaker's gaze (M = 49.72, Std.Error = 1.52) was different than the listener's gaze (M = 28.53, Std.Error = 1.18), [t = 12.928, p = 0.001]. No significant difference was found on the mutual gaze during the time before the referring expression [t = -1.421, p = 0.156].

There was also a significant difference on the proportional gaze of the speaker to the referent object (M = 49.72, Std.Error = 1.52), towards gaze on other objects (M = 33.07, Std.Error = 1.37), [t = 6.154, p = 0.001].

Finally, we compared the proportional gaze of the speaker to the referent object during the utterance and in the 1 second period before the utterance, however there was no significant difference, [t = -1.854, p = 0.064]. There was a difference however, as expected, on the listeners gaze during (M = 39.45, Std.Error = 1.11) and before (M = 28.5309, Std.Error = 1.18) the utterance, [t = 9.745, p = 0.001].

6.3. Eye gaze timing

Further, we investigated when each interlocutor turned their gaze to the object that was referred to verbally. In 450 out of the 582 cases the speaker was already looking at the object in a time window of 1 second prior to their verbal reference to it. On average during the 1s window they turned their gaze to the object 0.796s before uttering the referring expression [t = 83.157, p = 0.001], which is supported by the literature.

At least one of the listeners looked at the referent objects during the speaker's utterance in 537 out of the 582 cases. On average they started looking at the proximity area of the

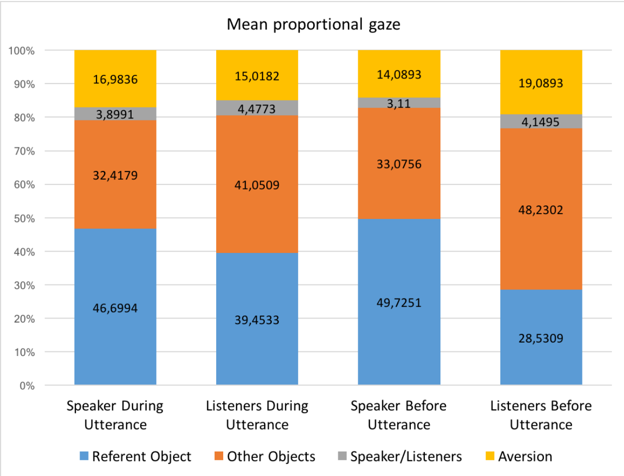

Figure 5: Mean proportional gaze per referring expression for speaker and listeners. a) The speaker's gaze during the time of the reference. b) The combined listeners' gaze during the speaker's referring expression. c) The speaker's gaze during the 1 second period before the reference. d) The combined listeners' gaze during the 1 second period before the speaker's referring expression.

referent object 344.5ms after the speaker started uttering a referring expression [$t = 19.355$, $p = 0.001$].

6.4. Mutual gaze and joint attention

We extracted all occasions where the group was jointly attending the area around the referent object, but also the occasions where either the speaker or the none of the listeners looked at the object: table 1.

N = 582	BU [-1..0]	DU [0..t]
None	58	21
Both	339	479
Speaker only	111	24
Listeners only	74	58

Table 1: Joint attention to the referent object area before (BU) and during (DU) speaker's references

During the preliminary analysis of the corpus we noticed that in many cases at least one of the listeners was already looking at the area around the referent object before the speaker would utter a referring expression (figure 5). It was our intuition that the salient object was already in the group's attention before. We looked at -1s before the utterance and automatically extracted these cases, as can be seen on table 2.

N = 582	BU [-1s]
None	163
Both	168
Speaker only	155
Listeners only	96

Table 2: Count of occasions where the referent object has already been in focus of attention (-1s before speaker's referring expression)

Finally we looked at the occasions where the interlocutors established mutual gaze. The table below shows cases where the speaker looked at one of the listeners or where at least one of the listeners looked at the speaker during and before referring expressions. Mutual gaze indicates where the speaker and one of the listeners look at each other, and as can be seen this is very rare during or before referring expressions.

N = 582	BU [-1..0]	DU [0..t]
Mutual gaze	10	20
Speaker at Listeners	52	72
Listeners at Speaker	80	143

Table 3: Count of occasions where the speaker looked at the listeners, the listeners looked at the speaker and mutual gaze during and before the speaker's referring expressions

7. Discussion

The presented corpus contains recordings of a variety of multimodal data, is processed and annotated and provides researchers with the possibility to explore multi-modal, multi-party turn-taking behaviours in situated interaction. While its strength lies in the rich annotation and the variation in conversational dynamics it also has some limitations.

One of the limitations is the granularity of eye-gaze annotation. Even though a high-end eye-gaze tracking system was used it was not always possible to disambiguate which of the potential objects were being gazed at. Given a visual angle threshold we are provided with a certain confidence measure defined by the angle difference to each object towards identifying the objects of attention. That limited our analysis to the area around the object rather than a precise measure of the object itself. Another limitation is that we did not analyse the object movements or the participants' pointing gestures. These might explain some of the visual attention cases prior to the verbal referring expressions.

Moreover, the use of the eye tracking glasses and motion capture equipment had the disadvantage that they were quite intrusive. Participants complained about fatigue at the end of the interactions and it was also qualitatively observed that their gaze-head movement coordination changed once wearing the glasses.

In most occasions the listeners and the speakers were looking at the area of the referent object before the referring expression was uttered which could potentially mean that the object was already on the group's attention. In some cases the listeners' visual attention was brought to the object area by referring language. Similarly to the referred literature this potentially shows that gaze has indications on the salient objects in a group's discussion and that can be used for reference resolution. It is also our intuition that objects that establish higher fixation density during referring expressions are considered to be more salient and can potentially resolve the references.

There are a few cases where neither the speaker nor the listener looked at the referent objects; in such cases text saliency algorithms (Evangelopoulos et al., 2009) or other multimodal cues such as pointing gestures (Lücking et al., 2015) could be combined to resolve the reference to the salient objects.

Typically speakers direct the listeners' attention using verbal and non-verbal cues and listeners often read the speaker's visual attention during referring expressions to get indication on the referent objects. As in literature we found that speakers and listeners gazed at each other a lot during the references to establish grounding on the referent objects. In very few cases however, they also established mutual gaze (looking at each other at the same time) during those references.

8. Conclusions

The current paper presents a corpus of multi-party situated interaction. It is fully transcribed and automatically annotated for eye-gaze, gestures and spoken language. Moreover, it features an automatic eye-gaze annotation method where the participant's gaze is resolved in 3d space; a visualisation tool is also used to qualitatively examine parts or the whole corpus in terms of conversational dynamics, turn taking and reference resolution. We annotated object references and investigated the proportional gaze from both the perspective of the speaker and the listeners. Finally, we quantitatively described the data of the corpus and gave further indications on how the corpus can be useful by the research community. Both the annotated corpus and visualisations are available at: `https://www.kth.se/profile/diko/page/material`.

9. Acknowledgements

The authors would like to thank Joseph Mendelson for moderating the sessions and helping with the recordings of this corpus. We would also like to acknowledge the support from the EU Horizon 2020 project BabyRobot (687831) and the Swedish Foundation for Strategic Research project FACT (GMT14-0082).

10. Bibliographical References

Admoni, H. and Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63.

Bohus, D. and Horvitz, E. (2009). Models for multiparty engagement in open-world dialog. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–234. Association for Computational Linguistics.

Bohus, D. and Horvitz, E. (2010). Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, page 5. ACM.

Bolt, R. A. (1980). *'Put-that-there': Voice and gesture at the graphics interface*, volume 14. ACM.

Borji, A. and Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207.

Bruce, N. D. and Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3):5–5.

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

Evangelopoulos, G., Zlatintsi, A., Skoumas, G., Rapantzikos, K., Potamianos, A., Maragos, P., and Avrithis, Y. (2009). Video event detection and summarization using audio, visual and text saliency. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3553–3556. IEEE.

Georgeton, L. and Meunier, C. (2015). Spontaneous speech production by dysarthric and healthy speakers: Temporal organisation and speaking rate. In *18th International Congress of Phonetic Sciences, Glasgow, UK, submitted*.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

Gross, S., Krenn, B., and Scheutz, M. (2017). The reliability of non-verbal cues for situated reference resolution and their interplay with language: implications for human robot interaction. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 189–196. ACM.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

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.

Kennington, C., Kousidis, S., and Schlangen, D. (2013). Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information. *Proceedings of SIGdial 2013*.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., and Rieser, H. (2010). The bielefeld speech and gesture alignment corpus (saga). In *LREC 2010 workshop: Multimodal corpora–advances in capturing, coding and analyzing multimodality*.

Lücking, A., Pfeiffer, T., and Rieser, H. (2015). Pointing and reference reconsidered. *Journal of Pragmatics*, 77:56–79.

Meena, R., Skantze, G., and Gustafson, J. (2014). Data-driven models for timing feedback responses in a map task dialogue system. *Computer Speech & Language*, 28(4):903–922.

Mehlmann, G., Häring, M., Janowski, K., Baur, T., Gebhard, P., and André, E. (2014). Exploring a model of gaze for grounding in multimodal hri. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 247–254. ACM.

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.

Prasov, Z. and Chai, J. Y. (2008). What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 20–29. ACM.

Prasov, Z. and Chai, J. Y. (2010). Fusing eye gaze with speech recognition hypotheses to resolve exophoric references in situated dialogue. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 471–481. Association for Computational Linguistics.

Rayner, K. (1995). Eye movements and cognitive processes in reading, visual search, and scene perception. In *Studies in visual information processing*, volume 6, pages 3–22. Elsevier.

Schlangen, D. and Skantze, G. (2009). A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 710–718. Association for Computational Linguistics.

Sheikhi, S. and Odobez, J.-M. (2012). Recognizing the visual focus of attention for human robot interaction. In *International Workshop on Human Behavior Understanding*, pages 99–112. Springer.

Skantze, G., Johansson, M., and Beskow, J. (2015). Exploring turn-taking cues in multi-party human-robot discussions about objects. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 67–74. ACM.

Staudte, M. and Crocker, M. W. (2011). Investigating joint attention mechanisms through spoken human–robot interaction. *Cognition*, 120(2):268–291.

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.

Sziklai, G. (1956). Some studies in the speed of visual perception. *IRE Transactions on Information Theory*, 2(3):125–128.

Winograd, T. (1972). Understanding natural language. *Cognitive psychology*, 3(1):1–191.