Comparison of Human-Human and Human-Robot Turn-Taking Behaviour in Multiparty Situated Interaction

Martin Johansson
KTH Speech Music and Hearing
Stockholm, Sweden
vhmj@kth.se

Gabriel Skantze
KTH Speech Music and Hearing
Stockholm, Sweden
gabriel@speech.kth.se

Joakim Gustafson
KTH Speech Music and Hearing
Stockholm, Sweden
jocke@speech.kth.se

ABSTRACT

In this paper, we present an experiment where two human subjects are given a team-building task to solve together with a robot. The setting requires that the speakers' attention is partly directed towards objects on the table between them, as well as to each other, in order to coordinate turn-taking. The symmetrical setup allows us to compare human-human and human-robot turn-taking behaviour in the same interactional setting. The analysis centres around the interlocutors' attention (as measured by head pose) and gap length between turns, depending on the pragmatic function of the utterances.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine System – Human Information Processing; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – Natural Language

Keywords

Situated dialogue; turn taking; multiparty human-robot dialogue

1. INTRODUCTION

Robots of the future are envisioned to help people perform tasks, not only as mere tools, but as autonomous agents interacting and solving problems together with humans. Such interaction will be characterised by two important features that need to be taken into account when modelling the spoken interaction. Firstly, joint problem solving is in many cases *situated*, which means that the spoken discourse will involve references to objects in the shared physical space. When speaking about objects, humans typically pay attention to these objects and gaze at them. To solve the task efficiently, interlocutors need to coordinate their attention, resulting in so-called joint attention [1]. Secondly, the robot should be able to solve problems together with several humans (and possibly other robots) at the same time, which means that we also need to model *multi-party* interaction. A central problem for spoken dialogue systems is turn-taking — i.e., to decide how to yield the turn and when to take the turn. In multi-party interaction, this becomes even more challenging. An obvious signal that humans use for yielding the turn in a face-to-face setting is to gaze at the next speaker. However, in situated interaction, where the gaze is also used to pay attention to the objects which are under discussion, it is not obvious how this shared resource is used.

In this paper we present an experimental setup where two human subjects are given a team-building task to solve together with a robot, as shown in Figure 1. The task is an adaption of the “Desert Survivor” team-building game [2]. The team is to collaboratively prioritize a set of items based on their usefulness for survival in a desert after a hypothetical plane crash. Cards with pictures of the items are placed on the table by which the speakers are seated, thereby constituting an area for joint attention. The robot can direct its attention, using head pose and eye movement, to objects on the table as well as to the human interlocutors. Since all three interlocutors are seated in an equilateral triangle pattern, this symmetrical setup allows us to compare human-human and human-robot turn-taking in the same interactional setting. Thus we can use this setup to explore three different questions at the same time:

1. How do humans behave when yielding and taking the turn between each other? This needs to be modelled in order for the robot to understand who has the floor, but also for the robot to employ a more human-like behaviour.
2. By comparing the subjects' behaviour towards each other with that towards the robot, we can investigate to what extent they interact with the robot as if it was a human interlocutor.
3. By comparing the robot's turn-taking behaviour with the subjects' behaviour, we can evaluate how human-like the robot's current behaviour is, and how it could be improved.

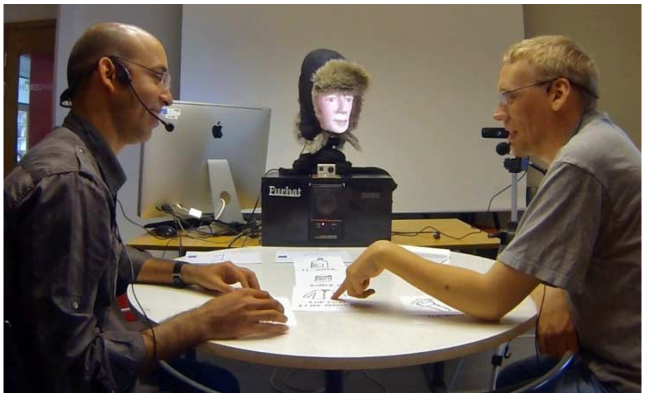

Figure 1. The human-human-robot interaction setup.

UM3I'14, November 16, 2014, Istanbul, Turkey

ACM 978-1-4503-0652-2/14/11…$15.00
http://dx.doi.org/10.1145/2666242.2666249

2. BACKGROUND

In spoken interaction, humans coordinate their turn-taking using several different signals, such as prosody, syntax and gaze. To yield the turn, speakers for example gaze at the interlocutor [3, 4], use syntactically complete phrases, and end the phrase with falling or rising pitch [5]. To keep the turn, they typically look away, use filled pauses, syntactically incomplete phrases, or a flat final pitch. There have been several attempts at building statistical or computational models of how turn-taking is coordinated (e.g., [6, 7, 8]). However, with a few exceptions (e.g., [9]), little work has been done on building such models for turn-taking in multiparty settings or situated interaction, where the interlocutors' attention to each other shares the same resources as their attention to the objects under discussion.

Multi-party interaction differs from dyadic interaction in several regards [10]. First, in a dyadic interaction there are only two different roles that the speakers can have: speaker and listener. In multiparty interaction, humans may take on many different roles, such as side participant, overhearer and bystander [11]. Second, in dyadic interaction, it is always clear who is to speak next at turn shifts. In multiparty interaction, this has to be coordinated somehow. The most obvious signal is to use gaze to select the next speaker [12].

However, in situated interaction, speakers also naturally look at the objects which are under discussion. The speaker's gaze can therefore be used by the listener as a cue to the speaker's current focus of attention. Speakers seem to be aware of this fact, since they naturally use deictic expressions accompanied by a glance towards the object that is being referred to [13]. In the same way, listeners naturally look at the referent during speech comprehension [14], and their gaze can therefore be used as a cue by the speaker to verify common ground. Thus, eye gaze acts as an important coordination device to achieve joint attention in situated interaction. This has been shown to clearly affect the extent to which humans otherwise gaze at each other to yield the turn. Argyle and Graham [15] studied dyadic interactions involving additional targets for visual attention. Objects relevant to the task at hand were found to attract visual attention at the expense of the other subject.

In order to model how a robot should be able to direct its attention in a human-like manner, and be able to understand human attentional behaviour, we need to study how humans coordinate turn-taking in situated multi-party interaction. However, we cannot assume that humans will automatically behave towards robots as they do towards other humans. Thus, we cannot exclusively rely on studies of human-human interaction.

In [16], we originally presented the experimental setup also described here, together with an initial analysis of how the subjects direct their attention during turn changes. In this paper, we extend that analysis in two ways. Firstly, we label the data depending on the pragmatic function of the utterances that constitute the turn, and carry out a more detailed analysis based on these labels. Secondly, we investigate the gap lengths between the turns.

3. METHOD

3.1 Experimental Setup

It is not trivial to utilize the subjects' gaze in a human-robot interactional setting. Gaze trackers can be very accurate, but they are also limited in field-of-view, or (if head worn) too invasive. In addition, they are not very robust to blinking or occlusion, and typically need calibration. In this study we instead rely on head pose tracking, which is a more simple and robust approach. This way, we will not be able to capture quick glances or track more precise gaze targets. However, previous studies have found head pose to be a fairly reliable indicator for gaze in multi-party interaction, given that the targets are clearly separated [17, 18, 19].

The experimental setup (as described in the Introduction) was designed around a round table at which the two human subjects and the robot were placed. Subjects were seated on static chairs at fixed locations to have the triad placed in an equilateral triangle pattern. This setup has the advantage that the predicted target areas for visual attention are as separated as possible in order to elicit head rotation instead of only eye gaze. Each subject was monitored by a Microsoft Kinect sensor, responsible for tracking the head pose of its subject as well as the angle to the most prominent audio source.

The robot in the experiment was the back-projected human-like robot head Furhat, which is capable of combining head pose and eye gaze to direct attention [20]. In controlled experiments on multi-party and situated interaction, it has been shown that subjects can infer the target of Furhat's gaze with a high accuracy [20, 21].

The gaze of the robot was automated, whereas the speech was controlled by a wizard. The wizard decided when the robot should say something, and what it should say, by selecting one of several predefined utterances from an interface on a networked computer. When speaking, the gaze was directed either towards a subject or, while speaking about an item, towards the table. When not speaking, the robot's gaze was directed towards the participant coinciding with the most prominent audio source detected by the Kinect sensors.

The adapted "Desert Survivor" exercise comprised three iterations of five unique items to discuss and rank, each iteration starting with the robot asking the subjects to open a numbered envelope containing the items. The robot could then, during the discussion that followed, express opinions about the relative importance of two items, say one of two predefined positive or negative things about an item, ask the subjects for their opinions, answer questions with a yes or a no or confess that it did not know. The goal was to make the robot behave in a similar way as an average human subject and not take on any specific role.

Eight subjects aged 23-41, divided into four pairs, participated in the data collection. The mean age was 30.5, with a standard deviation of 5.42. Three of the participants were female and five male. We recorded twelve iterations, three per pair of subjects. The recorded interactions, excluding instructions, lasted a total of 78 minutes with an average iteration length of 6.5 minutes (*SD* = 1.37).

3.2 Data Annotation

Each channel of the recorded audio was automatically segmented into Inter Pausal Units (IPUs) with a maximum of 500ms internal silence and then manually transcribed. The logged utterances from the robot were added as a third channel. A turn was defined as a sequence of one or more non-interrupted IPUs by the same speaker. A turn change was defined as two neighbouring turns by two different speakers, with a maximum overlap of 0.5 seconds and maximum gap of three seconds between the two turns. These thresholds were tuned in order to identify turns that were somewhat related to each other. Table 1 presents an example of a sequence of turn changes extracted from one of the dialogues.

We also wanted to make a more in-depth analysis of the subjects' turn-taking behaviour, depending on the pragmatic function of the constituting turns (i.e., their "dialogue act"). To this end, we used Amazon Mechanical Turk to assign labels to the two turns involved in each of the 682 detected turn changes. The utterances were presented as manually transcribed text and included the robot's utterances, without revealing if a specific utterance was made by a human or by the robot. The first turn was labelled based on its "forward-looking function" and the second turn on its "backward-looking function", inspired by the DAMSL coding scheme [22]. Thus, the same turn was in many cases annotated both in a forward-looking and a backward-looking context.

Table 1. Short example from a dialogue.

Speaker	Extracted Turn
Robot	yes I do think so ... I would say the pistol
Human 1	but I think this argument about being heard is quite strong actually.
Robot	yes
Human 1	and you might want to put the gun somewhere further up
Human 2	eh but water is more important than being heard and we do have the giant wreck which can be seen from the air
Human 1	okay

To make the annotation task simple, we restricted the number of labels to a minimum set of categories that we deemed interesting for the analysis. It should be noted that due to the automated turn change extraction, not all annotated turn changes involved two turns related to each other in a forward- or backward-looking context. Each turn change was labelled by three annotators, and the final label was based on the majority decision of the annotators. In total, there were 25 unique annotators, and majority decisions for all but 18 turns. For the yielding turns (Table 2), 75% of the labels were unanimous decisions, and 52% for the taking turns (Table 3), yielding a 3-annotator Fleiss' kappa of 0.619 and 0.463, respectively.

Table 2. Labels for yielded turns.

Label	N	Example(s)
Question	211	is the flashlight useful what do you mean
Non-question	453	the flashlight needs batteries yes / the flashlight is useful okay / I don't agree

Table 3. Labels for taken turns.

Label	N	Example(s)
Answer	146	yes / the flashlight is useful
(Dis-)Agreement	150	okay / I don't agree
Clarification request	87	what do you mean
Other	261	the flashlight needs batteries is the flashlight useful

For this analysis, we will focus on the three most common label pairs: *Non-question → Other* (36%), *Question → Answer* (26%), and *Non-question → Agreement* (22%). The latter will henceforth be referred to as *Statement → Agreement* based on inspection of the utterances involved.

4. RESULTS

For statistical analysis, we have used two-tailed tests and chosen an alpha level of 0.05.

4.1 Visual Focus of Attention

First we analysed where the turn yielder's attention was directed at turn shifts. To do this, we defined a window between 2 seconds before the end of the turn until the next speaker started to speak. Then we analysed whether the head pose was directed at the other interlocutors somewhere during this window. This resulted in four categories: attention towards both interlocutors, only towards the next speaker, only towards the other interlocutor (which did not take the turn), or towards none of the interlocutors (i.e., only at the table).

As seen in Figure 2, in a majority of the turn changes, the yielder does not give exclusive gaze towards the taker, which shows that head pose is not a very strict signal for turn-taking in this setting, and that it is generally quite open for both interlocutors to respond. A chi-squared test also shows that when a human yields the turn to the robot, it looks at the next speaker more often, compared to when yielding a turn to a human (χ^2= 9.25, dF=3, p=0.026).

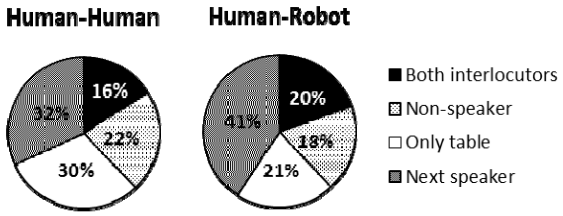

Figure 2. Distribution of the yielders' head pose at turn changes depending on taker.

Figure 3 breaks down these figures for the most common label pairs. In the human-human scenario, statements followed by an agreement appear to be directed to the next speaker, as opposed to questions followed by an answer where the targets are more varied. This is also in contrast to the human-robot case where the Question-Answer pair is the combination that appears to be most clearly addressed to the robot.

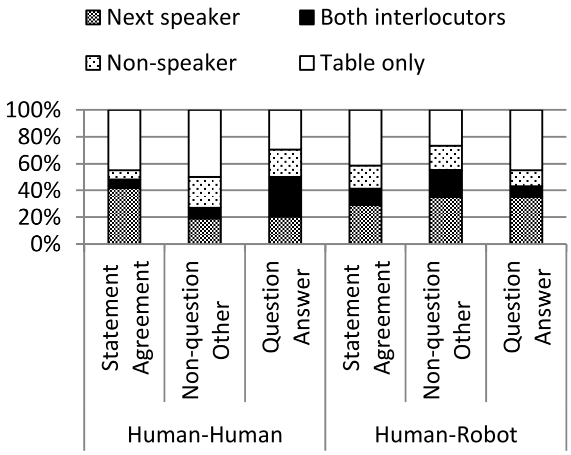

Figure 3. Distribution of the yielders' head pose for different label pairs

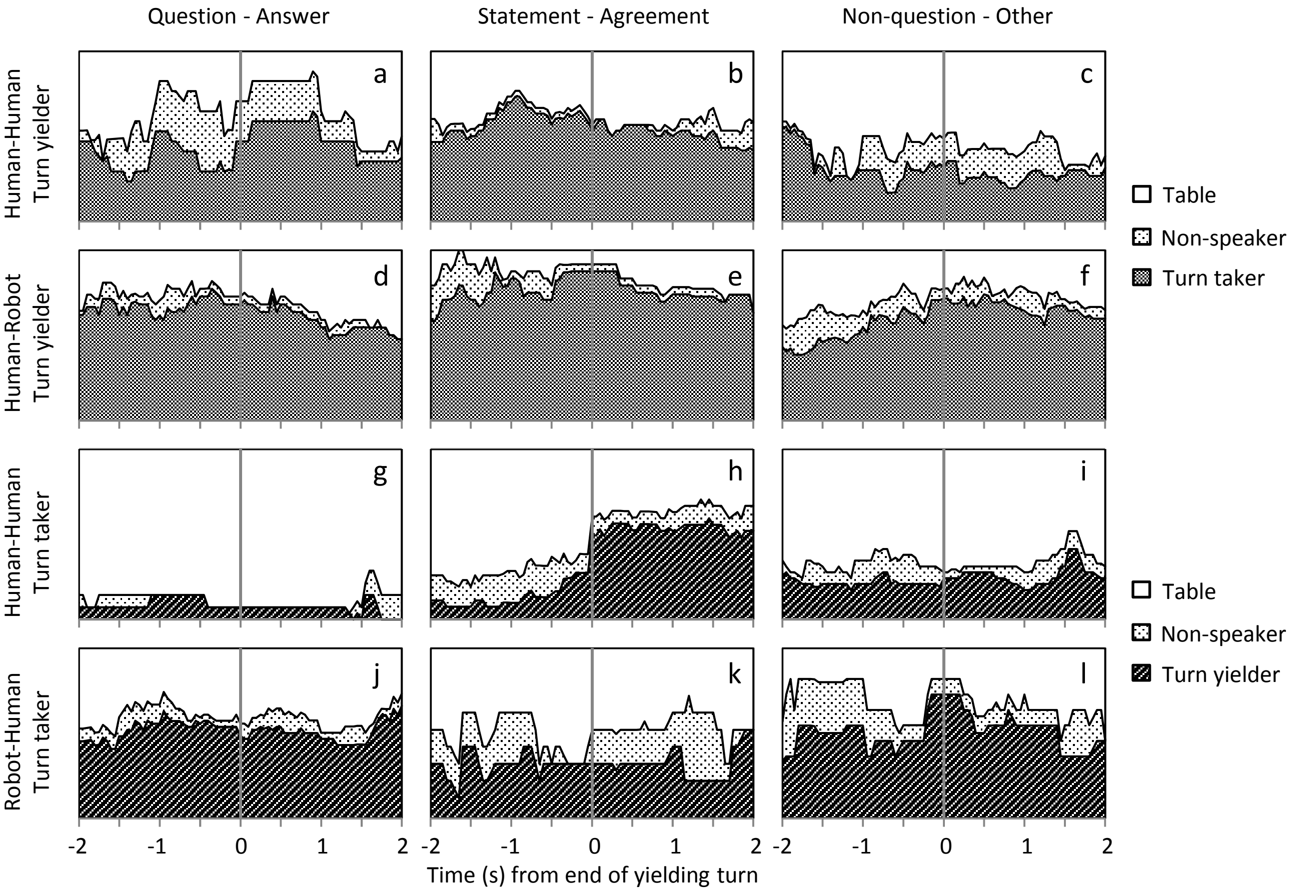

Figure 4. Distribution of human's visual focus of attention near turn changes.

The difference in proportions for the target of visual attention among different labels in a human-human turn change is significant, $\chi^2(6, N=204)=28.3722$, $p<0.001$. This is however not the case with human-robot turn changes, $\chi^2(6, N=203)=9.1038$, $p=0.168$.

This suggests that different pairs of pragmatic functions produce different head pose behaviour, and that the presented symmetrical setup can be used to capture such differences. The results also indicate that the robot is not triggering the same behaviour. This could be due to the robot's behaviour or due to the subjects' expectations of the system.

4.2 Target Distributions near End of Turns

Next, we made a more detailed analysis of the turn changes where the yielding speaker looked at the next speaker or both interlocutors, i.e. turn-shifts that were more clearly addressed. In Figure 4, the average distribution of both the yielder's and the taker's head pose are plotted in a window two seconds before and after the end of the yielding turn.

4.2.1 Question-Answer

For the *Question-Answer* pairs, the head pose patterns of the yielding human speaker differed both in shape and proportion depending on who the next speaker was. In the human-human turn changes (Figure 4a), the yielding speaker appears to have shifted visual attention to and away from the table and both interlocutors, whereas the visual attention in human-robot turn changes mostly focused on the robot (Figure 4d). The more consistent and exclusive attention given to the robot compared to the human can be interpreted as the robot mostly answering directed questions while humans to a larger extent also answer open questions. Explanations for this difference could be the wizard's behaviour when choosing to answer questions, or that the human interlocutors tried to address the robot more clearly. Similarly, the human answering a question asked by the robot looked at the robot to a high degree (Figure 4j), while a human answering a question from the other human mostly focused on the table (Figure 4g).

4.2.2 Statement-Agreement

In the *Statement-Agreement* pair, the yielding human speaker primarily looks at the next speaker and seldom at the other interlocutor (Figure 4be). This could be interpreted as a directed statement, where the speaker expects a reaction from a specific interlocutor.

The human taking the turn by providing an agreement is initially focused mostly on the table and then transfers attention to the yielding human speaker at the end of turn (Figure 4h). Thus, for agreement, mutual gaze seems to be important. This is in stark contrast to the corresponding Question-Answer pattern (Figure 4g). The pattern is also not present when the human expressed an agreement to a statement made by the robot (Figure 4k), once again suggesting that the robot is perceived differently.

4.2.3 Non-question-Other

For the third category, *Non-question-Other*, the distribution of attention is mostly flat in the human-human turn changes, both for the yielding speaker (Figure 4c) and for the turn taker (Figure 4i). This is not very surprising, since this category captures pairs of turns that are not as clearly related as the other categories. Once again, humans look more at the taker when it is the robot, as compared to when the other human takes the turn (Figure 4cf).

4.3 Gap between Turns

Next we wanted to explore the gap (response time) between the turns. We found the gap length to not be normally distributed, and therefore use median values and non-parametric tests. The median gaps for human-human and human-robot turn changes are shown in Figure 5.

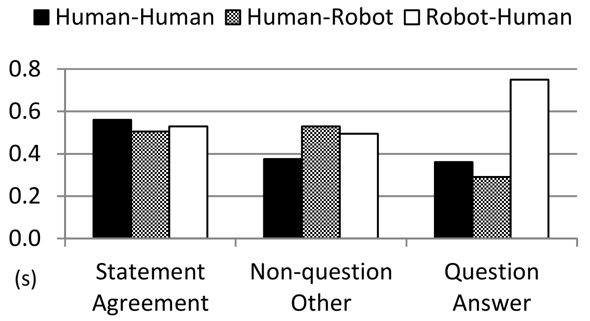

Figure 5. Median gap between turns.

The results indicate that the response time for wizard-controlled robot was quite close to a human in the same situation. However, a Kruskal-Wallis test showed that the three combinations of speakers had different gap lengths (χ^2=9.5593, dF=2, p=0.008). An adjusted follow-up multiple comparison test identified the Robot-Human as significantly different from the other two. Thus, the difference does not lie in the robot's response time, but rather in how the humans respond to the robot. The main difference seems to be in the Question-Answer exchanges. This could be due to the type of questions asked by the wizard, or due to the automated gaze not clearly signalling addressees when asking questions.

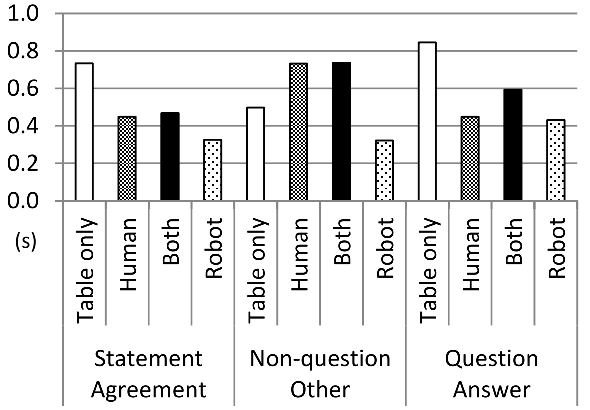

Figure 6. Breakdown of median gap length in human-human turn changes

Figure 6 contains a breakdown of the median gap between two turns in a human-human turn change based on the most common label pairs.

A Kruskal-Wallis test showed that the gap lengths are significantly different among the four classifications of visual focus of attention (χ^2=13.8919, dF=3, p=0.003). An adjusted follow-up multiple comparisons test identified the targeting of both interlocutors as significantly different from targeting only the robot. The short response time for the human taker when the yielder is looking at the robot is probably because the taker needs to start speaking faster to grab the turn.

5. CONCLUSIONS AND DISCUSSION

We can now revisit the three questions we posed in the Introduction and see to what extent the experimental setting presented here allows us to address these.

Firstly, simply by looking at the two humans' behaviour towards each other, it is clear that the attention is often directed towards the objects on the table and utterances are often not clearly directed towards a specific interlocutor. However, the focus of attention (as measured by head pose) is clearly different depending on the pragmatic function of the utterances that constitute the turn. At the end of questions, the speaker gaze more at both interlocutors, which means that they are typically not very directed. The interlocutor who answers the question pays more attention to the object under discussion, and very little attention to the person asking the question. This is quite different from statements followed by an agreement, where the speaker more clearly attends to one interlocutor, seeking agreement. The addressee, in turn, gazes back while giving the agreement. Looking at gap length, we can see that when an interlocutor wants to take the turn, but is not clearly attended by the turn yielder, the response time is much shorter, which indicates that the interlocutors monitor the current speaker's focus of attention in order to grab the turn.

The second question was whether the humans behave differently towards the robot, compared to how they behave towards each other. In general, they attend much more clearly to the addressee when yielding the turn to the robot, regardless of the type of exchange. The gaze is also much more stable, which is especially clear when looking at the Question-Answer category. Also, when taking the turn, humans look much more towards the turn yielder when it is a robot than when a human is yielding the turn (except for the Statement-Agreement category). Looking at gap length, we could also see that questions asked by the robot are not answered as quickly as questions asked by a human. These discrepancies indicate that humans to do not respond to the robot in exactly the same way as the other human when it comes to turn-taking. If the goal is not to create a fully human-like robot, this could of course be exploited, since it makes the detection of when the human is actually addressing the robot easier.

Regarding the third question, we can see that despite the Wizard-of-Oz setup, the robot manages to have a response time that is quite similar to the humans. The robot's automated gaze behaviour on the other hand differed in comparison to that of the humans. The robot's gaze was automated using readily available cues; if the robot said something about an item, its gaze was directed towards the table for the duration of the utterance, and in all other cases the robot directed its gaze towards the most prominent audio source. This is in stark contrast to the humans, who as listeners mostly looked at the table. In addition, the humans also exhibited different gaze behaviour for different dialogue acts, like looking at interlocutors when asking questions, or looking at the previous speaker when expressing agreement. This calls for a more sophisticated model for controlling the robot's gaze behaviour, if a more human-like behaviour is desired.

6. FUTURE WORK

The study presented here is just a first step towards building a model of situated multi-party interaction. The goal is to fully automate the robot's behaviour using the dialogue system framework IrisTK [23]. To this end, we will collect more data using the current setup, and then use machine learning to build models of where the focus of attention should be targeted, as well as how quickly the robot should respond, depending on the types of utterances exchanged.

We think that the symmetrical setup presented in this paper serves as an excellent test bed for evaluating such a model. By conducting analyses similar to the ones we have done here, we can measure to what extent the robot behaves similarly to the human interlocutors, and to what extent it triggers human responses which are similar towards the robot as those towards the other human.

7. ACKNOWLEDGMENTS

This work is supported by the Swedish research council (VR) project *Coordination of Attention and Turn-taking in Situated Interaction* (2013-1403).

8. REFERENCES

[1] Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In Joshi, A. K., Webber, B. L., & Sag, I. A. (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge, England: Cambridge University Press.

[2] Burgoon, J. K., Bonito, J. A., Bengtsson, B., Cederberg, C., Lundeberg, M., & Allspach, L. (2000). Interactivity in human-computer interaction: A study of credibility, understanding, and influence. *Computers in Human Behavior, 16*(6), 553-574.

[3] Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26*, 22-63.

[4] Oertel, C., Wlodarczak, M., Edlund, J., Wagner, P., & Gustafson, J. (2012). Gaze Patterns in Turn-Taking. In *Proc. of Interspeech 2012*. Portland, Oregon, US.

[5] Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology, 23*(2), 283-292.

[6] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech, 41*, 295-321.

[7] Morency, L. P., de Kok, I., & Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems, 20*(1), 70-84.

[8] Meena, R., Skantze, G., & Gustafson, J. (2013). A Data-driven Model for Timing Feedback in a Map Task Dialogue System. In *14th Annual Meeting of the Special Interest Group on Discourse and Dialogue - SIGdial* (pp. 375-383). Metz, France.

[9] Bohus, D., & Horvitz, E. (2011). Decisions about turns in multiparty conversation: from perception to action. In *ICMI '11 Proceedings of the 13th international conference on multimodal interfaces* (pp. 153-160).

[10] Traum, D., & Rickel, J. (2001). Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds. In *Proc. of IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems* (pp. 766-773). Seattle, WA, US.

[11] Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., & Ishiguro, H. (2012). Conversational Gaze Mechanisms for Humanlike Robots. *ACM Trans. Interact. Intell. Syst., 1*(2), 12:1-12:33.

[12] Vertegaal, R., Slagter, R., van der Veer, G., & Nijholt, A. (2001). Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of ACM Conf. on Human Factors in Computing Systems*.

[13] Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50*, 62-81.

[14] Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language, 38*(4), 419-439.

[15] Argyle, M., & Graham, J. A. (1976). The central Europe experiment: Looking at persons and looking at objects. *Environmental Psychology and Nonverbal Behavior, 1*(1), 6-16.

[16] Johansson, M., Skantze, G., & Gustafson, J. (2013). Head Pose Patterns in Multiparty Human-Robot Team-Building Interactions. In *International Conference on Social Robotics - ICSR 2013*. Bristol, UK.

[17] Katzenmaier, M., Stiefelhagen, R., Schultz, T., Rogina, I., & Waibel, A. (2004). Identifying the Addressee in Human-Human-Robot Interactions based on Head Pose and Speech. In *Proceedings of International Conference on Multimodal Interfaces ICMI 2004*. PA, USA: State College.

[18] Stiefelhagen, R., & Zhu, J. (2002). Head orientation and gaze direction in meetings. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (pp. 858-859).

[19] Ba, S. O., & Odobez, J-M. (2009). Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 39*(1), 16-33.

[20] Al Moubayed, S., Skantze, G., & Beskow, J. (2013). The Furhat Back-Projected Humanoid Head - Lip reading, Gaze and Multiparty Interaction. *International Journal of Humanoid Robotics, 10*(1).

[21] Skantze, G., Hjalmarsson, A., & Oertel, C. (in press). Turn-taking, Feedback and Joint Attention in Situated Human-Robot Interaction. *Speech Communication, 65*, 50-66.

[22] Allen, J. F., & Core, M. (1997). *Draft of DAMSL: Dialog act markup in several layers.* Unpublished manuscript.

[23] Skantze, G., & Al Moubayed, S. (2012). IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of ICMI.* Santa Monica, CA.