Visual Attention in Spoken Human-Robot Interaction

Maria Staudte
Department of Computational Linguistics
Saarland University
Saarbrücken, Germany
masta@coli.uni-saarland.de

Matthew W. Crocker
Department of Computational Linguistics
Saarland University
Saarbrücken, Germany
crocker@coli.uni-saarland.de

ABSTRACT

Psycholinguistic studies of situated language processing have revealed that gaze in the visual environment is tightly coupled with both spoken language comprehension and production. It has also been established that interlocutors monitor the gaze of their partners, a phenomenon called "joint attention", as a further means for facilitating mutual understanding. We hypothesise that human-robot interaction will benefit when the robot's language-related gaze behaviour is similar to that of people, potentially providing the user with valuable non-verbal information concerning the robot's intended message or the robot's successful understanding. We report findings from two eye-tracking experiments demonstrating (1) that human gaze is modulated by both the robot speech and gaze, and (2) that human comprehension of robot speech is improved when the robot's real-time gaze behaviour is similar to that of humans.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; J.4 [**Social and Behavioral Science**]: Psychology

General Terms

Experimentation, Human Factors, Measurement

Keywords

gaze, visual attention, experimental methods, user study

1. INTRODUCTION

Where people look is very closely coupled with what they hear and say. Psycholinguistic studies of situated language processing have revealed that speakers look at objects shortly before mentioning them, while listeners tend to look at mentioned objects in their visual environment shortly after hearing the reference. It has also been established that interlocutors monitor the gaze of their partners to establish "joint attention". Seeing what the partner looks at can provide valuable information about what is being talked about and further facilitate mutual understanding. We hypothesise that such gaze behaviour may also be beneficial in spoken human-robot interaction (HRI), and we present two eye-tracking experiments to evaluate this claim.

The close coupling of gaze with production has been established by several previous studies, e.g., [7]. It has been shown that referential gaze in speech production is part of the planning process for an intended utterance and typically precedes the onset of the corresponding linguistic reference by approximately 800msec - 1sec. [6, 14]. Further it has been established that listeners' visual attention is driven by what they hear ([12, 13, 19]). Among others, [1] have investigated exactly when people look at what they hear: people look approximately 200-300 msec after the onset of the referential noun at a suitable referent in their environment.

It has further been established that interlocutors monitor the gaze of their partners if they can (see [5] for a comprehensive account of joint attention). Studies investigating this kind of gaze in communication [8] have provided evidence that listeners use speakers' gaze to identify a target before the linguistic point of disambiguation (i.e., the point in the sentence where other possible interpretations can be eliminated and the sentence can be verified). They show that the speaker's gaze helps to identify possible referents of an utterance, even when it was initially misleading due to the experimental setup. Subjects can establish a mapping of the speaker's gaze to their own visual scene and, thus, still make use of the speaker's gaze early during comprehension.

Combining the results described above, we can envisage the following scenario: Two people (A and B) are talking about an object (e.g. a mug) that is visible to both of them. According to the gaze production pattern, A says "Pass me the mug, please." and looks at the mug approx. 1 sec before saying "mug". To confirm the heard information, listener B then looks at the mug approx. 300 msec after A started saying "mug". Taking the duration of the actual word "mug" into account, these patterns results in a 1,5 -2 sec time span between the speaker's gaze towards the mug and the listener's gaze to that same object. If additionally A and B can see each other, joint attention can be established throughout this communication. Listener B can follow A's gaze towards the mug right away and anticipate A's utterance about the mug. The time span between A's and B's gaze towards the mug is shortened dramatically and B can faster understand A's utterance. Furthermore, in a situation when there are several mugs, gaze may provide a means of referential disambiguation.

The above mentioned findings illustrate how gaze during spoken communication is systematically and automatically coupled to situated speech. For that reason, speakers can reliably monitor listeners' eye movements to see whether they have been understood.

HRI'09, March 11–13, 2009, La Jolla, California, USA.

Similarly, listeners can interpret speakers' eye movements to help rapidly identify, and disambiguate among, intended referents.

Considerable work has already been done on robot gaze in HRI, e.g., for turn-taking [4] or with respect to information structure of the generated utterance [15]. It was further established that the perception of robot gaze is coupled to the robot's head orientation by [9]. It has also been shown that robot head movement towards the speaker and away from the speaker can signal engagement in a conversation [17]. Moreover, it has been shown that robot gaze alternating between the listener and an object of interest at relevant dialogue points results in greater non-verbal engagement of the participants [21]. However, the on-line psycholinguistic findings from studies of human speech and gaze that have motivated our work have, to our knowledge, not yet been applied in HRI.

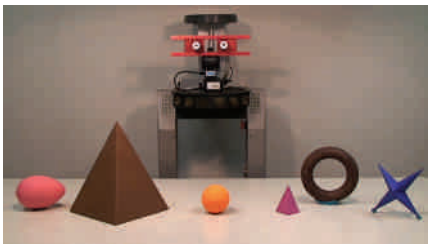
We hypothesise that people exploit robot gaze to comprehend a robot's utterances about its environment, and as a consequence seek to establish joint attention with the robot. We further hypothesise that humans integrate this visual information about gaze direction during language comprehension in a similar manner as in human-human communication. Specifically, we predict that the robot's gaze directly influences where people look in a scene (Prediction 1) and that this affects people's comprehension of the robot's utterance (Prediction 2).

The particular setting of the experiments is as follows. We recorded videos of a robot that looked at objects presented on a table in front of it while it produced statements about this scene. Our participants are eye-tracked while observing these videos. They are also instructed to determine the 'correctness' of the robot's statement with respect to the scene and to respond by pressing a button accordingly. We examine the human behaviour in response to the robot's gaze behaviour and/or errors in the statements' propositional truth. A pilot study [18] has already demonstrated this general method to be suitable for our purposes.

We consider two dependent measures: We use eye-tracking to monitor when people look at what and for how long; We also record response times of the participants in response to the robot's statement. While common evaluation methods in HRI (like questionnaires) often rely on personal interpretations of the users, we decided to make use of the traditional measures used in human-human studies, i.e., eye-tracking and response times. These measures have the advantage of being taken "on-line" at a sampling rate of 2 msec. That is, we observe the human behaviour during processing and we can directly relate it to the unfolding visual and linguistic events in the experimental stimuli.

Moreover, our design has several advantages over previous user or evaluation studies in HRI. The video-based presentation enables us to create a larger number of stimuli off-line and show them on-line. This is a precondition for collecting statistically reliable data. Furthermore, the off-line stimuli preparation allows us to carefully control and manipulate robot utterance production and the related robot gaze behaviour separately. Specifically the robot's gaze and speech can be synchronised to be similar to that of humans. In the presented studies, we made use of such gaze patterns for producing referential robot gaze, i.e., fixations towards an object approximately 1 sec before it is mentioned. Human visual attention can then be observed in response to both the robot utterance and gaze. Thus, we can measure the effect of robot gaze versus robot utterance on the participant's visual attention towards potential referents in the scene. More precisely, if robot gaze is indeed considered to be an expression of robot attention (and, hence, is potentially beneficial for communication), then we expect to observe that participants exploit this early piece of information to visually ground and fully understand the uttered sentence.

(a) Unambiguous condition.



(b) Ambiguous condition.

Original sentence: "Die Kugel ist neben einer Pyramide."
(Translation: "The sphere is next to a pyramid.")

Figure 1: Sample scenes from Experiment 1.

Although it might be argued that this is not true interaction, it has been shown that a video-based scenario without true interaction yields similar results to a live-scenario and can be considered to provide (almost) equally valuable insights into the subject's perception and opinion [20]. Further, the subjective perception of remote versus collocated agents (for both robots and virtual agents) has been studied by [11] and similar results were presented.

Using the experimental paradigm outlined above, we conducted two experiments. Experiment 1 examines whether human gaze is influenced by both robot speech (revealed by the listener's looks towards a mentioned object) and gaze (looks towards an object fixated by the robot). Experiment 2 examines the benefit of robot gaze for comprehension of robot speech. We compare human behaviour in response to videos in which robot gaze is correct, infelicitous or absent.

2. EXPERIMENT 1

2.1 Goal

In this study we investigate whether people's gaze is influenced by robot gaze and speech on-line. Participants saw the robot while it gave a description of several objects in its view. A description such as "The sphere is next to a pyramid." is accompanied by robot gaze to a sphere and then to a pyramid, each occurring shortly before the robot utters the corresponding noun phrases (Figure 1). This within-subjects design has one factor (ambiguity) with two levels. In one condition, the video shows among other shapes one sphere and one pyramid. In the second, ambiguous condition, there are two pyramids in the scene, both matching the utterance "The sphere is next to a pyramid." Both conditions require a positive answer since the statements of both conditions are always true.

Since participants need to verify the statement against the scene, we assume that their gaze behaviour is influenced by the robot's utterance. It is unclear, however, whether participants follow the robot's gaze as well. In the unambiguous condition, both robot gaze and speech refer to a unique target object. In the ambiguous condition, the robot's utterance identifies two potential referents

(two pyramids) while robot gaze is directed only towards the target pyramid. We observe and compare our participants' looks towards the target pyramid and the distractor object in both conditions to establish whether people follow robot gaze.

2.2 Methods

2.2.1 Participants

Forty-eight native speakers of German, mainly students enrolled at Saarland University, took part in this study (14 males, 34 females). Most of them had no experience with robots. They were told that the eye-tracker camera was monitoring their eye movements and pupil size to measure the cognitive load of the task on them.

2.2.2 Materials

A set of 16 items was used. Each item appeared in both conditions. One condition comprises a scene that is uniquely described by the uttered sentence. The other condition comprises a scene which is ambiguously described by the corresponding sentence. The ambiguity results from two potential target objects in the scene as shown in Figure 1.

We created 1920×1080 resolution video-clips showing a PeopleBot robot[1] onto which a pan-tilt-unit is mounted. This pan-tilt-unit carries a stereo camera which appears as the head and/or eyes of the robot. Note, that head orientation and eye-gaze of the robot is therefore identical.[2] The robot stands behind a table with a set of coloured objects in front of it. The objects are plain geometrical shapes of different colours and sizes. In the unambiguous condition (Figure 1(a)), each shape occurs only once on the table and the uttered sentence has a unique interpretation with respect to the scene. In the ambiguous condition (Figure 1(b)), two objects of the same shape (but of different colours and sizes) are target and distractor referents in a corresponding sentence. The video-clips each show a sequence of camera-movements consecutively towards the object mentioned first and the target object on the table. At the same time, a synthesised sentence of the form given in Example (1) is played back.

The robot fixations and the spoken sentence are timed such that a 'fixation' towards an object happens approximately one second prior to the onset of the referring noun phrase which is consistent with psychological findings about the co-occurrence of referential gaze and referring expressions in human speech production [7]. Because of these distinct time windows we can study both types of reactive human gaze separately: one being elicited by robot gaze (joint attention), the other being utterance-mediated (inspecting mentioned objects).

In both conditions the participant has to give a positive answer since both statements are true. Further, across the 16 items we balanced the stimuli with respect to target size (eight target objects are big and have small distractors and vice versa) and target location. In addition to the 16 item videos described above, we constructed 56 filler videos (of which 24 videos were used as items in Experiment 2).

[1]kindly provided by the DFKI CoSy/CogX group: http://www.dfki.de/cosy/www/index.html.

[2]Previous studies support the assumption that listeners use mostly head orientation as indicator for visual attention rather than eye-gaze itself so that a distinct realisation of the two does neither seem necessary nor is it technically possible at this stage (see [9] for HRI and [8] for HHI)

2.2.3 Procedure and Task

An EyeLink II head-mounted eye-tracker monitored participants' eye movements at a sampling rate of 500 Hz. The video clips were presented on a 24-inch colour monitor. Viewing was binocular, although only the dominant eye was tracked, and participants' head movements were unrestricted. For each trial, a video was played until the participant pressed a button or until an overall duration of 12 seconds was reached. There were two buttons side by side, one for each response option. The button configuration was chosen such that participants always had to use their main hand to press the "correct" button. After a drift correction interlude the next video clip was presented. The participants were instructed by a short text to attend to the scene and quickly decide whether the robot's statement was right or wrong with respect to the scene. To make the task appear more natural, participants were further told that their results were used as feedback in a machine learning procedure for the robot. The entire experiment lasted approximately 30 minutes.

2.2.4 Analysis

The presented videos are segmented into Interest Areas (IA), i.e., each video contains regions that are labelled "target" and "distractor". The output of the eye-tracker is mapped onto these IAs to yield the number of participant fixations on an object. The spoken utterance is a sentence similar to the one shown in Figure 1, describing the relation between a couple of objects. For our analysis the "pyramid" is encoded as the **target** reference. In the unambiguous condition, the "pyramid" refers to exactly one target object. In the ambiguous condition, the "pyramid" may refer to the **target** object *or* the **distractor** object since there are two pyramids in the scene.

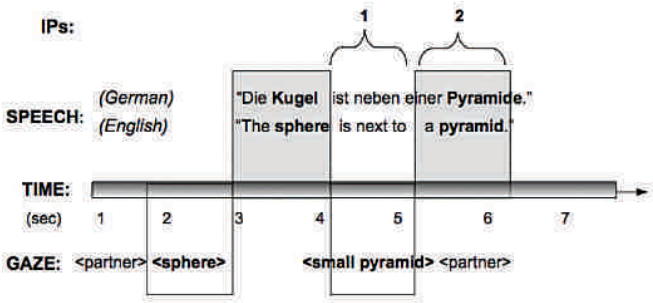


Figure 2: The approximate timing of utterance-driven robot gaze for the given sentence.

We segment the video/speech stream into two Interest Periods (IP) based on the onsets and offsets of the encoded linguistic events. The IPs identify the time regions when the robot head fixates the target object and when it refers linguistically to the target object (see Figure 2). For the analysis of the participants' fixations, we compute proportions of fixations per IA within each IP in a condition (fixations on an IA are divided by all fixations in this IP, i.e. proportions between 0 and 1). For each IP in particular, we compare the fixation proportions on the target and the distractor area between all conditions. IP1 is defined as the 1000 msec period preceding the onset of the target phrase, and contains the robot's fixation on the target object as well as some verbal content preceding the target noun phrase ("next to"). IP2 stretches from the noun phrase onset (including the determiner) to offset and has a mean duration of 674 msec (min=488, max=972 msec).

The offset of IP2 also marks the end of the sentence. The elapsed time between this offset and the moment of the button press com-

prises the response time.[3] For the statistical analysis of both the response time and the fixation proportions, we use the repeated-measures analysis of variance (ANOVA). Subject and item means are entered separately into the analyses. The fixation proportions factor IA (target, distractor) and condition (unambiguous, ambiguous).

2.2.5 Predictions

If robot gaze is not used, we expect participants to solely rely on the robot's utterance and thus fixate the distractor object more often in the ambiguous condition than in the unambiguous condition. If, however, participants do follow gaze, we expect to observe looks towards the target even before it is being mentioned (in IP1) because the robot's gaze precedes the target mentioning. Furthermore, if people interpret gaze as identifying the intended referent, they should continue to favour the target over the distractor when it is mentioned (IP2), even in the ambiguous condition. Since both conditions are true, and gaze is consistent with human behaviour, we expect to observe no difference in response times between both conditions. Indeed, a difference in response times would suggest that people were unable to use gaze effectively in the ambiguous condition.

2.3 Results

Fixations

We observed that participants look significantly more often at the target than at the distractor in both conditions. That is, there is a main effect for factor IA during IP1 ($F_1(1, 45) = 58.28$ and $F_2(1, 14) = 189.66$, with $p_1 < 0.005$ and $p_2 < 0.005$) and during IP2 ($F_1(1, 45) = 87.93$ and $F_2(1, 14) = 43.36$, with $p_1 < 0.005$ and $p_2 < 0.005$). That this effect is observable in IP1 indicates that participants in fact follow the robot's gaze towards the target object. Moreover, participants looked equally often at the target object in the ambiguous and the unambiguous condition as depicted in Figure 3, suggesting that participants followed robot gaze to the target even when there was another potential referent in the scene.

Further, we did not observe a main effect for ambiguity in either IP, i.e., the influence of an additional referent is not significant with respect to participants' gaze. In IP2, however, we found that participants looked more often towards the distractor object in the ambiguous condition than in the unambiguous condition. This effect may have caused the observed interaction effect between the factors IA and ambiguity in IP2 ($F_1(1, 45) = 9.19$ and $F_2(1, 14) = 5.68$, with $p_1 < 0.005$ and $p_2 < 0.05$). The increase in the number of looks towards the distractor object suggests that participants do notice the referential ambiguity and accordingly fixate the distractor in the ambiguous condition. Nevertheless, there is a strong preference for fixating the target object in both conditions which indicates that participants easily identify the target despite the scene ambiguity.

Response Times

As expected, we observe no significant difference in the response times ($F_1(1, 47) = 0.747$ and $F_2(1, 14) = 0.026$).[4] In both conditions participants are equally fast in determining the validity of the statement. The findings on both response time and the recorded eye movement data are coherent with our hypothesis that humans seek to establish joint attention with the robot, i.e. that they follow the robot's gaze to the target (Prediction 1). However, sentences in Experiment 1 were referentially ambiguous, possibly emphasizing the role of robot gaze. To reliably test Prediction 2, the influence of robot gaze when accompanying uniquely identifiable sentences needs to be explored which was done in Experiment 2.

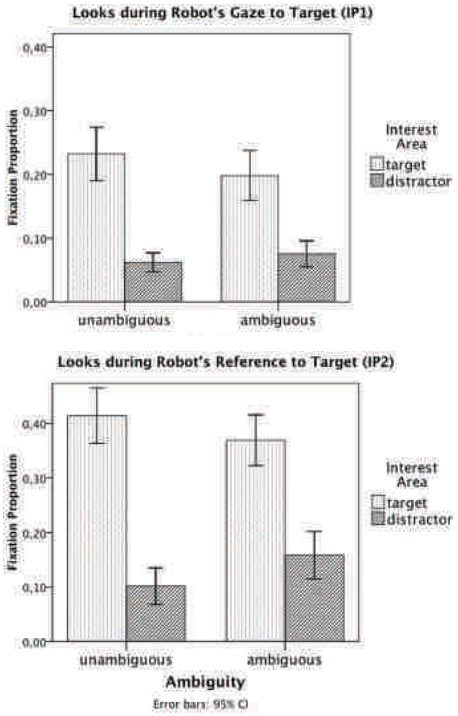


Figure 3: Fixation proportions by condition and interest area, for both interest periods.

3. EXPERIMENT 2

3.1 Purpose

Experiment 1 demonstrates that human gaze is influenced by both robot gaze and speech. In Experiment 2, we sought to further investigate the actual benefit of robot gaze. To separate the influence of robot gaze and speech we manipulate the *congruency* of our robot's gaze as a cue for intended meaning and the *validity* of the statements.

More precisely, in a 2×3 within-subjects design, we manipulate two factors: Statement validity (true or false) and gaze congruency. The latter denotes the match of the visual reference (established by the robot's gaze) with the linguistic reference (made in the robot's statement), and comprises three levels (congruent, incongruent, no robot gaze). We consider gaze to be congruent (and informative) when it is directed towards the same object that is going to be mentioned shortly afterwards (reference match) while it is considered as incongruent when gaze is directed to an object different from the mentioned referent (mismatch). In the third congruency level robot gaze is absent to provide a baseline condition in which the participants' visual attention is purely a response to the produced utterance. In Experiment 2, the robot's statement is of the form that is given in the example sentence below.

[3]Trials were excluded from this analysis if the participant gave a wrong answer. Wrong and correct button presses cannot be compared with respect to response times. Four percent of the trials had to be excluded for this reason.

[4]F_1 and F_2 yield the results of analyses based on subject means and item means, respectively.

Example:
"Der Zylinder ist groesser als die Pyramide, die pink ist."
("The cylinder is bigger than the pyramid that is pink.")

The scene provides two potential referents (e.g. two pyramids of different sizes and colours) one of which the robot mentions. One referent matches the description of the scene while the other does not, which determines the statement truth. The manipulation of both factors, statement validity and congruency, results in six conditions per item. Below, we provide an example for all conditions sentence (2) can appear in (given a corresponding scene depicted in Figure 4) :
Conditions for the example sentence given above:

1. True statement: "The cylinder is bigger than the pyramid that is pink."
   (a) Congruent (looks to mentioned object that makes sentence valid = (small) pink pyramid),
   (b) Incongruent (looks to another object that would make sentence invalid = (big) brown pyramid),
   (c) No robot gaze
2. False statement: "The cylinder is bigger than the pyramid that is brown."
   (a) Congruent (looks to mentioned object that makes sentence invalid = (big) brown pyramid),
   (b) Incongruent (looks to another object that would make sentence valid = (small) pink pyramid),
   (c) No robot gaze

3.2 Methods

3.2.1 Participants and Procedure

This study was run simultaneously with the first experiment. The items of one experiment were used as filler items for the other. Therefore, the participants as well as the procedure were identical for both experiments.

3.2.2 Materials

A set of 24 items was used. Each item consists of three different videos and two different sentences, i.e., appears in six conditions. Additionally we counterbalance each item by reversing the comparative adjective, i.e., from "bigger" to "smaller" such that the target becomes the distractor and vice versa. We obtain a total of twelve videos per item while ensuring that target size, location and colour were balanced. All versions show the same scene and only differ with respect to where the robot looks and whether it refers to the correct (target) object. All twelve object shapes appear twice as target-distractor pairs. The actual objects were pre-tested in order to make sure that their size and colour differences were easily recognisable. The questionnaire we used showed photographs of the original scenes excluding the robot. Twenty participants had to judge whether a given item sentence accurately described what was visible in the scene. The results suggest that object comparisons are easily assessed.

The videos were of the same type as in Experiment 1. The robot fixations and the spoken sentence are again timed such that a fixation towards an object happens approximately one second prior to the onset of the referring noun. In this experiment, we can observe the two types of reactive human visual attention in separate conditions: In addition to the fact that robot-gaze occurs in a time window preceding the uttered reference, we introduce a baseline condition not showing any robot-gaze at all. Since Experiments 1 and 2 were run simultaneously, we had 48 fillers (16 items from Experiment 1 and an additional set of 32 filler videos) for 24 item videos.

Twelve lists of stimuli each containing 72 videos were created. Each participant saw only one condition of an item and, in total, four videos in each condition. The order of the item trials was randomised for each participant individually.

3.2.3 Analysis

The IAs in this study contain the target and the distractor objects. The "pyramid" from the example sentence above is the **target** reference that has two referents in the scene when it is mentioned: the small, pink **target** pyramid *or* the large, brown **distractor** pyramid.

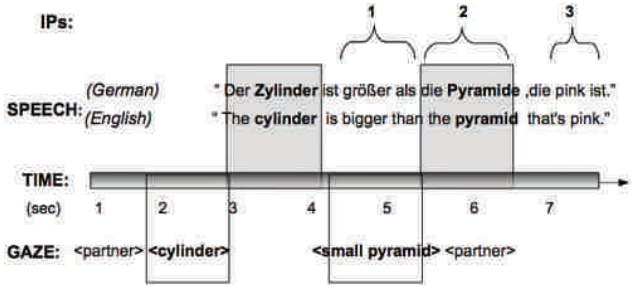


Figure 5: The approximate timing of utterance-driven robot gaze, in condition *true-congruent*.

We segmented the video/speech stream into three Interest Periods (IP) as depicted in Figure 5. IP1 is defined as the 1000 msec period ending at the onset of the target phrase (IP2). It contains the robot's fixation towards the target object as well as some verbal content preceding the target noun phrase (e.g. "bigger than"). IP2 stretches from the target phrase onset (including the determiner) to offset and has a mean duration of 674 msec. IP3 is defined as the 700 msec period beginning at the onset of the disambiguating colour adjective. For the analysis of the participants' fixations, we compute proportions of fixations per IA within each IP in a condition (as described for Experiment 1). For each IP individually, we compare the fixation proportions on the target and the distractor area between all conditions.

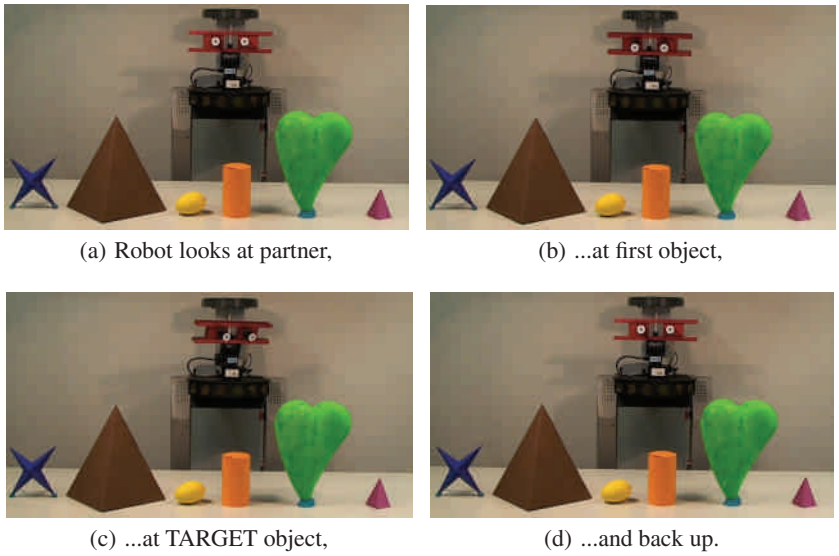
The adjective denoting the colour of the referent completes the linguistic reference and identifies the actual target. Only at that point in time is it possible to judge the statement validity, which is why it is called the linguistic point of disambiguation (LPoD).[5] The elapsed time between this adjective onset and the moment of the button press is therefore considered as the response time.

As in Experiment 1, the respective means are entered separately for subjects and items into the Repeated-Measures ANOVA. Both, response time means and fixation proportions, are analysed with two factors: statement validity and robot gaze congruency.

3.2.4 Predictions

In Experiment 1 we found that people exploit robot gaze to resolve a reference. In Experiment 2, we can compare between the presence and absence of robot gaze and the congruency of the latter in order to evaluate the utility of robot gaze. Based on our hypothesis and supporting findings from Experiment 1, we expect participants' gaze to be mediated by robot speech. We particularly expect

[5] A similar design, also featuring late linguistic disambiguation with early visual disambiguation by means of gaze-following, was already successfully tested in a study on human-human interaction by [8] .



(a) Robot looks at partner, (b) ...at first object,

(c) ...at TARGET object, (d) ...and back up.

Figure 4: Sample scene from Experiment 2.

to observe this when robot gaze is absent since participants' fixations within the scene should then be driven by the robot's utterance.

Since our item sentences reveal the target object, i.e., which object is being mentioned, only at the end of a sentence, participants can keep several hypotheses about referents until the LPoD. We therefore expect gaze throughout the utterance to reveal the listener's hypothesis about the intended referent. Specifically, in IP1 we predict gaze-mediated fixations while in IP2 we expect fixations on both the target and distractor IAs when the robot mentions the target noun phrase. Based on where the robot looks and what it says, we expect participants to preferably fixate the IA that they *consider* to be the actual target. IP3 reveals the match (congruent condition) or mismatch (incongruent condition) of visual and linguistic references made by the robot. Since the statement (and therefore the linguistic reference) has to be judged for its validity, we expect participants to then preferably fixate the *actual* target IA.

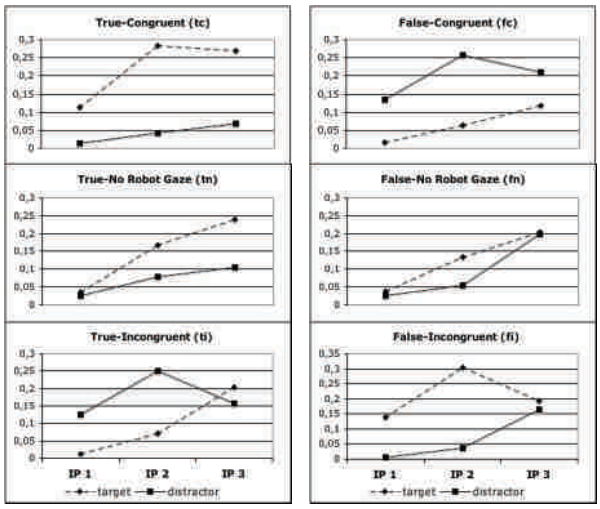
For response times, a main effect of statement validity is expected due to the bias in our stimuli (true statements have faster response times than false statements). We also expect a main effect of gaze congruency: If participants exploit robot gaze, they can anticipate the validity of statements in those stimuli when gaze is congruent with the statement. In contrast, when gaze is incongruent with the statement, we expect that participants anticipate a proposition that eventually does not match with the actual robot statement. Hence, we assume slower response times for incongruent robot gaze. Since the absence of gaze neither facilitates nor complicates the judgement of the statement validity, we predict intermediate response times for this condition.

3.3 Results

Fixations

In Figure 6 we have plotted the average fixation proportions of our participants on the IAs (target and distractor) within each IP.[6] On the left-hand side, the *true* conditions are depicted. In all three of these graphs, the robot utters the same sentence about the target

[6]Differences reported here as significant were statistically significant in pairwise post-hoc comparisons.



(a) Fixation proportions for true statements. (b) Fixation proportions for false statements.

Figure 6: Fixation proportion means for all three interest periods.

(e.g. the small pink pyramid from the example sentence) while only its gaze behaviour differs. When comparing these three graphs, the impact of the presence or absence of robot gaze on the participants' fixations becomes evident:

IP1: During IP1, robot gaze is the only potential cue to the intended target (e.g. big or small pyramid). The upper left graph depicts the *true-congruent* condition (tc), i.e., the robot's gaze refers to the same object that the (true) statement refers to, namely the target. This graph also shows significantly more fixations on the target IA than the distractor IA. The middle-left graph plots fixations in the *true-no robot gaze* condition (tn). In contrast to the conditions containing robot gaze, there are almost no fixations on the target or distractor IAs during this IP. The bottom-left graph depicts fixations in the *true-incongruent* condition (ti), when the robot looks at

the distractor IA. Similarly (yet reversed), we observe significantly more fixations on the distractor IA than on the target IA.

IP2: In IP2 the robot utters the target noun phrase (e.g. "the pyramid"). The fixation pattern observed in IP1 for both gaze conditions is enhanced in IP2. The mentioning of the noun phrase increased fixations on the already preferred IA. In the absence of gaze, participants begin to fixate the small target pyramid which satisfies the linguistic description so far ("The cylinder is bigger than the pyramid").

IP3: This IP contains the LPoD specifying which pyramid is indeed being mentioned. In the tc-condition the robot gaze and statement match and so participants remain fixating the target IA and hardly look at the distractor. However, in the ti-condition the mismatch between visual and linguistic reference becomes apparent and participants have to realise that the robot's statement is not about the distractor object. Probably in order to re-judge the statement's validity, they start to look at the target IA as well (no significant difference between target and distractor IA now).

The fixation pattern is almost identical for false statements. What is being fixated by the robot, and therefore by the participant, is reversed. The statement is false in these conditions, i.e., the robot mentions the distractor object (e.g. big brown pyramid). *False - congruent gaze* therefore means that the robot also looks at the distractor object. Consequently, in the fc-condition participants mainly fixate the distractor IA as well. Note, that for both tn- and fn-conditions the videos are identical up to IP2. The fixation patterns nicely confirm this by showing a simultaneous fixation increase on the target IA in both conditions up to that point. The fixation patterns then diverges in IP3 according to the LPoD. In the no-gaze conditions, it becomes apparent that participants predict a suitable object as the referent (target) based on the available linguistic material. When robot gaze is present, however, it overrides this linguistic prediction: in the false-congruent condition, gaze-following to the distractor is observed even though the distractor does not fulfill the linguistic description given up to IP2.

Response Times

We found main effects for both statement validity and gaze congruency in the response times as plotted in Figure 7. Specifically, participants were significantly faster (at an average of 139.73 msec) when they had to give a positive answer than when the statement of the robot was false ($F_1(1,47) = 17.69$ and $F_2(1,23) = 7.93$, with $p_1 < 0.005$ and $p_2 < 0.05$). Gaze congruency also has a significant effect on response times ($F_1(2,46) = 13.55$ and $F_2(2,46) = 25.7$, with $p_1 < 0.005$ and $p_2 < 0.005$). In the absence of an interaction of the two factors, we can compare the three levels of the congruency factor independent of statement validity. In the congruent condition, i.e., when the robot looks towards the object that it is going to mention, participants are significantly faster (135.49 msec on average) in giving their response than when there is no robot gaze involved. Participants are faster in the *no robot gaze* condition than when the robot's gaze is incongruent with its statement (145.37 msec on average). The result is a cascaded response time pattern: true < false, congruent < no gaze < incongruent. A post-hoc pairwise comparison with a Bonferroni adjustment further reveals pairwise significant differences between response times in the *true-congruent* and *false-congruent* conditions ($F_1(1,47) = 11.45$ and $F_2(1,23) = 6.41$, with $p_1 < 0.005$ and $p_2 < 0.05$) and between the *true-no robot gaze* and *false-no robot gaze* condition ($F_1(1,47) = 6.14$ and $F_2(1,23) = 4.98$, with $p_1 < 0.05$ and $p_2 < 0.05$). The two *incongruent* conditions do not differ significantly with respect to response time.

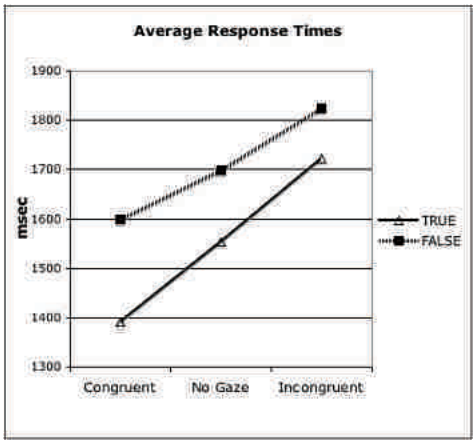


Figure 7: Average response times for true and false statements, per gaze congruency condition.

The response times clearly suggest that congruent gaze benefits and incongruent gaze disrupts comprehension relative to the *no robot gaze* condition (Prediction 2). This suggests that participants do associate robot gaze with the robot's statement about objects in the scene. We assume that they start building hypotheses about the statement's validity according to the robot's gaze and before the LPoD. Hence, participants are able to make their decisions faster when gaze is congruent with the statement than in those cases where there was no robot gaze in the video. On the other hand, when the robot's gaze is incongruent and leads the participant to a referent different from the mentioned one, the participant has to reassess the statement's validity and construct a new hypothesis. We suggest that this additional comprehension time occurring at the LPoD causes the slow-down in response time.

Concluding our results for Experiment 2, we find that the response time results support the interpretation of our findings from the observed eye movements described above and, similarly, suggest that participants follow both robot gaze and robot speech.

4. CONCLUSIONS AND FUTURE WORK

We have shown that detailed insights from situated human communication can be applied to human-robot-interaction. The presented evidence shows that this cognitively motivated robot-gaze behaviour is beneficial in HRI and that humans react in a manner typical of HHI to both robot speech and robot gaze.

More precisely, we predicted that the robot's gaze directly influences where people look in a scene (Prediction 1) and, further, that this affects people's comprehension of the robot's utterance (Prediction 2). The two studies presented in this paper revealed that participants make use of robot gaze, i.e., they follow it when it is available, which clearly supports Prediction 1. This is the case even when the task does not require them to do so: In our experiments it suffices to match the uttered statement against the scene without paying attention to the robot's movements. In Experiment 2 in particular, we showed that robot gaze which is congruent with the uttered sentence helps human interlocutors to faster judge utterances than if robot gaze was absent. On the other hand, when robot gaze was incongruent with the utterance, i.e., it referred to a different object, it slowed people down. This evidence clearly confirms Prediction 2.

We have further shown that humans integrate robot gaze online during incremental utterance comprehension and that this influences comprehension speed. We suggest that this effect is due to a reduction (congruent gaze) or increase (incongruent gaze) of the hypothesis space during comprehension, as a result of cues provided by cognitively motivated robot gaze behaviour.

The impact of these findings for the design of systems controlling robot gaze is considerable. We conclude that referential robot gaze contributes to a faster and more fluent communication and thus is to be preferred over a robot that does not look at the objects it is talking about. However, when the robot is not entirely certain about the location of a referent (or which object to look at) it is advisable not to initiate any fixations, since these may disrupt the comprehension of the user.

Moreover, we suggest that the proposed experimental design is genereally suitable to investigate what beliefs humans have about robots and their capabilities. The attribution of beliefs, goals and desires to others is a crucial skill in social interaction ([2, 3]). This capability is necessary in order to realise, for instance, what the interaction partner is attending to and why. Attention, intentions and beliefs are important aspects of human-robot interaction as well. Of course, a robot is not expected to act like a human, but with increased communicational skills the expectations towards the robot will also rise.

Previous research has considered a Theory of Mind (ToM) model and its utility for human-robot interaction from the robot perspective. Scasselatti and colleagues, for instance, attempted to implement two ToM models on a robot system [16]. Their aim was to equip a robot with a system that enables the robot to "engage in natural human social dynamics" by maintaining a ToM for the human partners it interacts with. Others have attempted to investigate what mental models people have for robots [10] with a focus on the appearance of the robot and the anthropomorphism that people associate with it. With our design, however, we can investigate more precisely what features of a ToM humans build when interacting with a robot. For instance, what do people think about the robot's cognitive capabilities. Is the robot considered to have (visual) attention which reflects internal (and even intentional) states? Which modality do people preferably trust in and consider more reliable? A ToM model that is based on HRI instead of HHI might be simpler and yet more effective when applied to a robot system.

The studies conducted so far and reported here provide support for our hypothesis that people pay attention to robot gaze, exploit it and integrate the gained information during utterance comprehension. We conclude that humans consider robot gaze to be meaningful and that cognitively motivated gaze behaviour can therefore contribute to more natural and fluent HRI in general.

5. ACKNOWLEDGMENTS

The research reported of in this paper was supported by IRTG 715 "Language Technology and Cognitive Systems" funded by the DFG. Many thanks to Afra Alishahi for giving valuable feedback.

6. REFERENCES

[1] G. Altmann and Y. Kamide. Now you see it, now you don't: Mediating the mapping between language and the visual world. In J. Henderson and F. Ferreira, editors, *The Interface of Language, Vision, and Action: Eye Movements and The Visual World*, pages 347–386. Psychology Press, NY, 2004.

[2] S. Baron-Cohen, D. Baldwin, and M. Crowson. Do Children with Autism Use the Speaker's Direction of Gaze Strategy to Crack the Code of Language? *Child Development*, 68:48–57, 1997.

[3] S. Baron-Cohen, A. Leslie, and U. Frith. Does the autistic child have a "theory of mind"? *Cognition*, 21:37–46, 1985.

[4] J. Cassell, O. Torres, and S. Prevost. Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation. *Machine Conversations*, pages 143–154, 1999.

[5] P. D. Chris Moore, Philip J. Dunham, editor. *Joint Attention Its Origins and Role in Development*. LEA, 1995.

[6] Z. M. Griffin. Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, 82:B1–B14, 2001.

[7] Z. M. Griffin and K. Bock. What the eyes say about speaking. *Psychological Science*, 11:274–279, 2000.

[8] J. Hanna and S. Brennan. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57:596–615, 2007.

[9] M. Imai, T. Kanda, T. Ono, H. Ishiguro, and K. Mase. Robot mediated round table: Analysis of the effect of robot's gaze. In *Proc. of 11th IEEE ROMAN '02*, pages 411–416, 2002.

[10] S. Kiesler and J. Goetz. Mental models of robotic assistants. In *Conference on Human Factors in Computing Systems*, pages 576–577, 2002.

[11] S. Kiesler, A. Powers, S. Fussell, and C. Torrey. Anthropomorphic interactions with a robot and robotÐlike agent. *Social Cognition*, 26:169–181, 2008.

[12] P. Knoeferle and M. W. Crocker. The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science*, 30:481–529, 2006.

[13] P. Knoeferle and M. W. Crocker. The influence of recent scene events on spoken comprehension: evidence from eye-movements. *Journal of Memory and Language (Special issue: Language-Vision Interaction)*, 57:519–543, 2007.

[14] A. Meyer, A. Sleiderink, and W. Levelt. Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, 66:B25–B33, 1998.

[15] B. Mutlu, J. Hodgins, and J. Forlizzi. A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior. In *Proceedings 2006 IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS'06)*, Genova, Italy, 2006.

[16] B. Scassellati. Theory of mind for a humanoid robot. In *1st IEEE/RSJ International Conference on Humanoid Robotics (Humanoids 2000)*, Cambridge, MA., 2000.

[17] C. L. Sidner, C. Lee, C. Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164, 2005.

[18] M. Staudte and M. W. Crocker. The utility of gaze in human-robot interaction. In *Proceedings of "Metrics for Human-Robot Interaction", Workshop at ACM/IEEE HRI 2008*, Amsterdam, Netherlands, 2008.

[19] M. K. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, and J. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634, 1995.

[20] S. Woods, M. Walters, K. L. Koay, and K. Dautenhahn. Comparing Human Robot Interaction Scenarios Using Live and Video Based Methods: Towards a Novel Methodological Approach. In *Proc. AMC'06, The 9th International Workshop on Advanced Motion Control*, 2006.

[21] A. Yamazaki, K. Yamazaki, Y. Kuno, M. Burdelski, M. Kawashima, and H. Kuzuoka. Precision timing in human-robot interaction: Coordination of head movement and utterance. In *Proceedings of CHI '08*, 2008.