The effect of robot gaze on processing robot utterances

Maria Staudte & Matthew W. Crocker
Department of Computational Linguistics
Saarland University
Saarbrücken, Germany
{masta, crocker}@coli.uni-saarland.de

Abstract

Gaze during situated language production and comprehension is tightly coupled with the unfolding speech stream in a manner that provides interlocutors with relevant on-line information about both what we intend to say and what we have understood. This paper investigates whether people similarly exploit such gaze when listening to a robot make statements about the shared visual environment. On the basis of two eye-tracking experiments exploiting different tasks, we report evidence (a) that people's visual attention is influenced on-line by both the robot's gaze and speech, (b) that congruent gaze (to mentioned objects) facilitates comprehension, and (c) that robot gaze does indeed influence what listeners think the robot intended. This supports the view that spoken interaction with artificial agents such as robots benefits when those agents exhibit cognitively-derived real-time speech-mediated attention behaviour.

Introduction

Situated spoken language processing has been widely and carefully studied, mainly within the Visual World Paradigm. On one hand, *listeners* fixate potential referents in visual scenes as soon as there is enough linguistic information to delimit a set of potential referents (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Among others, Altmann and Kamide (2004) have shown that people look at the referent about 200-300ms after the onset of the referential noun. On the other hand, *speakers* look at what they plan to talk about: Referential gaze in speech production typically precedes the onset of the corresponding linguistic reference by about 800ms-1sec (Griffin, 2001; Meyer, Sleiderink, & Levelt, 1998).

While these findings robustly characterise speech-mediated gaze to objects and events in visual scenes, listeners' visual attention may also be influenced by speakers' gaze. When people communicate about a common scene they establish mutual gaze as well as joint attention to support fast and robust understanding (see Moore (1995)). Hanna & Brennan (2007), for instance, have shown that listeners use speakers' gaze to identify a visual referent even before the linguistic point of disambiguation uniquely identifies the mentioned referent. While there exist some significant results on robot gaze and its role for engagement in human-robot interaction (HRI; see Staudte & Crocker (2009) for a literature review and initial results), we investigate by means of objective measures whether HRI would benefit from robot speech and gaze production that is as closely coupled as in humans.

In order to better control experimental conditions and to obtain statistically relevant data, we chose a video-based setting with a tele-present robot. Although it might be argued that this is not true interaction, it has been shown that a tele-present robot has similar effects on the subjects' perception and opinion as a physically present robot (Kiesler, Powers, Fussell, & Torrey, 2008; Woods, Walters, Koay, & Dautenhahn, 2006).

We hypothesize that if people integrate cognitively motivated robot gaze with the actual robot utterance, they combine both, gaze and utterance, into one reference resolution mechanism. That is, we predict that people will follow robot gaze to objects in the scene (Prediction 1) and that robot gaze leads people to make assumptions about what the intended referents are (Prediction 2). We present results from two experiments supporting the hypothesis that people exploit cognitively motivated robot gaze during utterance comprehension. Specifically, we show both, that people follow robot gaze to an object when its gaze is available and that this affects the resolution of the linguistic reference.

Experiment 1: Integration of concurrent visual and linguistic information

Experiment 1 investigates the on-line influence of robot gaze on situated utterance comprehension. The presence of the robot as speaker and its gaze direction provide a *visual reference*: In addition to verbally referring to an object the robot can also use gaze as a visual pointer to an object. To separate the influence of robot gaze and speech we manipulate two factors in a 2×3 within-subjects design: *Statement validity* (true/false) and *gaze congruency*. Congruency denotes the match of the visual reference established by the robot's gaze and the linguistic reference 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, providing a baseline condition: Participants' visual attention is driven purely by the utterance. The robot's statements were produced by the Mary TTS system (Schroeder & Trouvain, 2001) and are of the form given below.

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 for the final noun phrase (e.g. two pyramids of different sizes and colours) one

Spoken Reference to:	Gaze Reference to: Target (small pyramid)	Competitor (big pyramid)	none
Target (small pyr.)	true congruent	true incongruent	true no gaze
Competitor (big pyr.)	false incongruent	false congruent	false no gaze

Table 1: 2x3 design of both experiments.

of which the robot mentions. While the small pink pyramid (the target) matches the example description of the scene depicted in Figure 1, the big brown pyramid (the competitor) does not. This way the mentioned colour of the target determines the statement truth. The manipulation of both factors - statement validity and congruency - results in six conditions per item. An sample of all conditions given the example sentence and scene is provided in Table 1.

The comparison of the two conditions (gaze/no-gaze) reveals when and how precisely robot gaze changes participant behaviour, while the congruency factor allows us to observe whether people use robot gaze as an early cue to mentioned referents and whether this can influence comprehension.

Methods

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.

Materials A set of 24 items was used so each participant saw four different items in each of the six conditions shown in Table 1. Each item consists of three different videos crossed with two different sentences. Additionally we counterbalance each item by reversing the comparative adjective, i.e., from "bigger" to "smaller" such that targets become competitors and vice versa. We obtain a total of twelve video/utterance pairs 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 which object it refers to. The objects are all plain geometric shapes that were pre-tested to make sure that their size and colour differences were easily recognisable. We used 48 fillers for 24 item videos such that participants saw a total of 72 videos. The robot's gaze and the spoken sentence are timed such that it looks towards an object approximately one second prior to the onset of the referring noun.

Task & Procedure Participants were instructed 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. An EyeLink II head-mounted eye-tracker

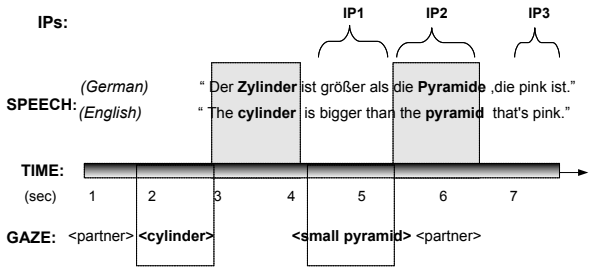


Figure 2: The approximate timing of utterance-driven robot gaze, in condition *true-congruent*.

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 and the next video started. Participants always had to use their dominant hand to press the "correct" button. The experiment lasted 30 minutes.

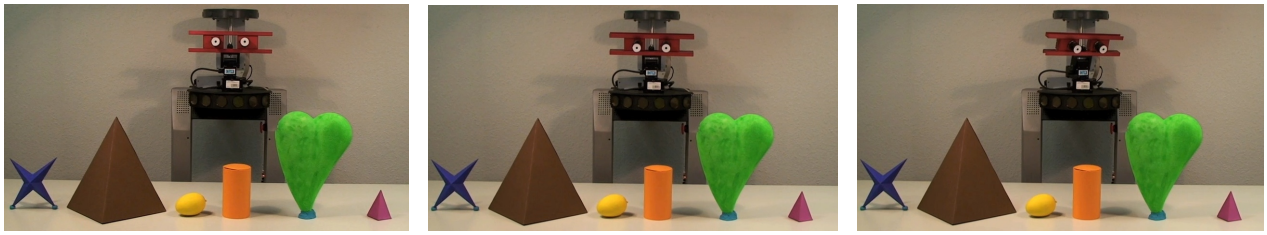
Analysis The presented videos are segmented into Interest Areas (IA) labelled, for instance, *target* or *competitor*, with eye-tracker output mapped onto these IAs to yield the number of participant fixations on an IA. The spoken utterance (see Example sentence) describes the relation between the central object, the *anchor*, and the target or competitor. The noun "pyramid" from the example sentence is a partial linguistic reference that has two referents in the scene: the small, pink pyramid as target *or* the large, brown pyramid as competitor.

We segmented the video/speech stream into three Interest Periods (IP) as depicted in Figure 2. IP1 is defined as the 1000ms period ending at the onset of the target noun (IP2). It contains the robot's gaze towards the target object as well as verbal content preceding the target noun phrase (e.g. "bigger than the"). IP2 stretches from the target noun onset to offset (mean duration of 674ms). IP3 is defined as the 700ms period beginning at the onset of the disambiguating colour adjective. Consecutive fixations within one IA were pooled as one inspection. We compute proportions of inspections per IA within an IP in a condition (summed for each IA across trials and divided by the total number of inspections in this IP). For each IP, we compare the inspection proportions on the target and on the competitor IA across 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).[1] The elapsed time between this adjective onset and the moment of the button press is therefore considered as response time (RT).

Our sentences include three different comparisons (size,

[1] 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 Hanna and Brennan (2007) .



(a) Robot looks at partner, (b) then at the anchor, (c) and at TARGET object.

Figure 1: Sample scene from the experiments.

height and width) and two complementary versions of each comparative, e.g. "bigger than" and "smaller than", in order to counterbalance target and competitor objects within one scene. We coded the comparative type (bigger, smaller) and the distractor IAs (small object for 'bigger than' and big object for 'smaller than') accordingly. When the partial sentence "The cylinder is bigger than" has been uttered, i.e. before the target noun is mentioned, at least two objects in the scene match this partial description: the target and a small distractor next to the anchor. Inspections that occur in this time region (IP1) therefore reveal whether people anticipate one, both or no potential referents at this stage. To conduct this extended analysis, we also coded the distractor objects as IAs. In the next time region (IP2), the target noun is uttered which extends the utterance to "The cylinder is bigger than the pyramid". While "pyramid" identifies two possible referents, only the target object is compatible with the comparative.

Predictions *IP1* ("The cylinder is BIGGER THAN THE"): In the no-gaze conditions, we expect the tendency to fixate objects that are smaller than the cylinder, i.e., the pink target pyramid and the small distractor. When gaze is present, we expect people to look more at the object fixated by the robot.

IP2 ("The cylinder is bigger than the PYRAMID"): In the absence of gaze there are two pyramids of which only the pink target pyramid is a suitable referent for the incomplete linguistic description. Therefore, if people process the robot utterance similarly to human utterances, we expect people to clearly prefer fixating the target pyramid over the competitor. However, when robot gaze is present and it is considered relevant, incongruent gaze is expected to interrupt this fixation behaviour.

IP3 ("The cylinder is bigger than the pyramid that's PINK."): While robot gaze occurs earlier in the sentence, in IP1, the factor determining sentence truth (LPoD) occurs at the end of the sentence, in IP3. That is, if gaze has a lasting effect on reference resolution, we expect to observe that people continuously fixate the visually salient referent when the linguistic reference is consistent with it. We expect incongruency to elicit fixations on both potential referents.

With respect to RT, we predict that participants respond faster to true statements than to false ones, mainly because the "correct" button is pressed by their strong hand and because verification is typically faster than falsification. More

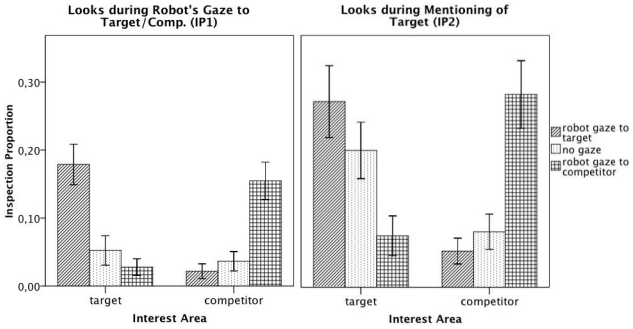


Figure 3: Inspection proportions in three gaze conditions and IP1 and IP2.

interestingly, we 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, we expect that participants anticipate a proposition that eventually does not match the actual statement. Hence, we assume slower RT for incongruent robot behaviour.

Results

The initial results of Experiment 1, including target and competitor inspections as well as RT, have been presented and explained in more detail in Staudte and Crocker (2009) and are reported below for the sake of convenience.

Since sentence truth does not play a role in IP1 and IP2 (because the LPoD only occurs in IP3), we collapsed each two conditions where trials are identical up to IP2. That is, conditions true-congruent and false-incongruent are collapsed into the condition "gaze to target", true-incongruent and false-congruent are collapsed into "gaze to competitor" and the two no-gaze conditions are merged into "no gaze".

In IP1, when the robot looks towards either the target or the competitor, we observe that people clearly follow this gaze and inspect the according IA.[2] We also observe a tendency for participants to look at the target more often than at the competitor in the no-gaze condition although this is not statistically significant. In IP2, when the robot mentions the target noun, the main effect of robot gaze and the interaction

[2]Main effect of robot gaze: $F(2,94) = 21.96, p < 0.001$ and significant interaction between gaze direction and IA.

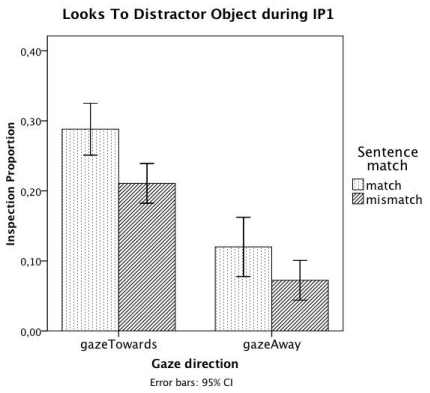


Figure 4: Inspection proportions on distractor in two gaze conditions (towards or away from distr.) for both comparative versions.

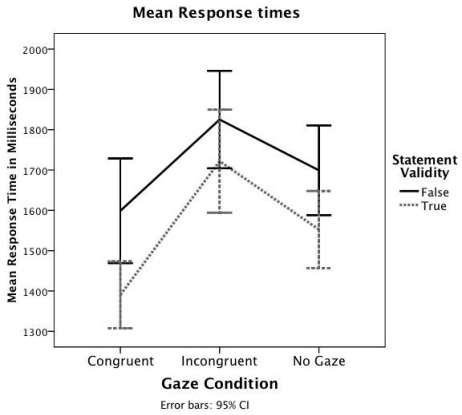


Figure 5: Average response times (RT) for all six conditions.

effect remain. Moreover, when the target noun is uttered and no robot gaze is available, people inspect the target IA significantly more often than the competitor IA ($F(1,46) = 15.53$ with < 0.001) despite the fact that the uttered sentence is still ambiguous at that point (see Figure 3).

We also look at the effect of robot gaze on a more fine-grained level of utterance processing: Based on the comparative in the sentence, people prefer the target over the competitor since it is the most probable referent. The presence of robot gaze, however, introduces additional (and potentially conflicting) information since it draws attention to either target or competitor. Thus, when gaze is available, preference for the object that fulfills the partial utterance is no longer observable, instead participants follow the robot's gaze.

We conducted a similar analysis (as for target and competitor) for both distractor objects next to the anchor. For the no-gaze conditions, the analyses clearly show that people inspect the distractor IA that matches the uttered comparative significantly more often, which confirms previous findings from human-human studies. Since the size difference of the two distractors does not seem to have an effect on the predictive power of the comparative, we collapsed the two IAs for the analysis of the gaze-conditions and obtain the following factors: Sentence match (distractor either matches or mismatches the comparative) and gaze direction (is or is not in the general direction of robot gaze, i.e., when the robot looks at the target/competitor located further away, its gaze passes one of the distractors). The results shown in Figure 4, suggest that the direction of the robot's gaze is indeed a very dominant cue that mainly determines where people look. Nevertheless, we find main effects for both robot gaze direction ($F(1,46) = 71.0, p > 0.001$) and the comparative match ($F(1,46) = 9.77, p > 0.001$).

In IP3, we analyse the IAs, target and competitor, separately, each with respect to the original factors statement validity and gaze congruency.We observe a main effect of statement validity ($F_{targ}(1,47) = 14.81, p > 0.001$ and $F_{comp}(1,47) = 38.7, p > 0.001$) suggesting that people look significantly more often at the object with the actually mentioned colour. Moreover, we observe a main effect of gaze congruency ($F_{targ}(2,94) = 3.42, p > 0.05$ and $F_{comp}(2,94) = 03.01, p > 0.06$) and a robust interaction effect of the two factors ($F_{targ}(2,94) = 4.05, p > 0.05$ and $F_{comp}(2,94) = 4.94, p > 0.1$). The interaction suggests that in congruent conditions people continuously inspect the object fixated and mentioned by the robot (which is the target in true statements and the competitor in false statements) whereas in incongruent conditions, subjects make a visual attention shift from the object fixated by the robot to the object actually mentioned by the robot.

The RT results confirm what the fixation data suggests. We observe main effects of both statement validity ($F(1,47) = 24.83, p < 0.001$) and gaze congruency ($F(2,94) = 17.2, p < 0.001$) (see Figure 5). Obviously, participants' visual attention is influenced by where the robot looks and consequently people respond faster when they already fixated an object that the robot mentions (congruent). When no gaze or, even worse, incongruent gaze is available, the required attention shift results in slower responses.

These RT results suggest that robot gaze as visual attention guidance influences (the speed of) utterance comprehension. Specifically, it was shown that linguistic reference is processed faster when people's visual attention is already on the according linguistic referent (Prediction 1). Whether robot gaze (as visual reference) further influences reference resolution and therefore the complete utterance representation is explored in Experiment 2.

Experiment 2: The effect of visual reference on linguistic reference resolution

The results of Experiment 1 suggest that robot gaze is a strong cue which guides visual attention in an automated way and that this influences comprehension. However, there are two possible explanations for the observed RT effects. Either robot gaze is a purely visual cue that influences visual attention but does not convey referential meaning or it is believed to express some kind of intentionality which elicits predictions about the intended referent of the speaker. If, indeed,

Condition	Gaze to:	Linguistic cue to: Comparative	Colour
false - no gaze:	—	Target	Competitor
false - congruent:	Target	Target	Competitor
false - incongruent:	Competitor	Target	Competitor

Table 2: Linguistic and visual cues to objects in three congruency conditions for a false sentence

the effect of robot gaze on people's visual attention is due to its referential meaning, we predict that robot gaze not only affects how fast references are resolved but also which object is believed to be the intended referent of the utterance.

Experiment 2 investigates how precisely robot gaze affects the resolution of the robot's intended referent when people have to correct the robot utterance. Specifically, we observe participants in response to a false robot utterance that is accompanied by either incongruent, congruent or no robot gaze.

Methods

Participants Thirty-six native speakers of German, again mainly students enrolled at Saarland University, took part in this study (12 males, 24 females).

Task & Procedure Participants were instructed to give a corrected sentence of the robot's utterance when they thought that the robot had made a mistake. They were further told to start the sentence with the same object reference that the robot started with, making it easier for the system to learn from the corrected sentences. The experiment was self-paced, i.e., participants decided when to continue to the next video by pressing a button. The procedure of Experiment 2 is otherwise identical to Experiment 1.

Materials There are two cues in the robot utterance identifying the correct referent. The first cue is the comparative (*bigger than* or *smaller than*) and the second cue is the object colour. False statements are false when these two cues do not identify the same referent, e.g., when the cylinder is not *bigger* than the *brown* pyramid. Thus, people can repair this utterance by changing either the comparative or the colour adjective. Details on referential variation for each condition are shown in Table 2.

A sentence produced by a participant consequently describes either the visually more salient object (based on robot gaze) or the alternative object when it seems linguistically more salient to them. Note that two factors constitute linguistic saliency, the comparative and the colour, and that their preference is measurable in the no-gaze condition: Any bias towards the comparative or towards the colour as an identifying feature is picked up and can be used as a baseline for the analysis of the two gaze conditions.

Analysis Produced corrections were recorded from the start of a video until the participant pressed a button to continue. For the analysis of the corrections, we annotated the sentences with respect to which object was described. The four categories assigned were Target, Competitor, No correction given (accepted as correct), Else (participants described a different object). We then computed proportions of corrections across trials (corrections, e.g., for the target, in one condition is divided by the total of trials in this condition). We use repeated-measures analyses of variance (ANOVA) with two factors (described object, gaze condition) for the statistical analysis of correction proportions.[3]

Predictions As previous research has shown, people prefer to use absolute (shape and colour) to relative features (size, location) for the production of referring expression (Beun & Cremers, 1998). We therefore expect to observe a general repair preference for the competitor, that is, for the object that is identified by colour. If visual attention is directed towards an object but no expectations with respect to the linguistic reference are elicited, we expect that people's repair pattern in the gaze-conditions will not differ significantly from the no-gaze condition. However, if robot gaze elicits expectations about the intended referent, we expect to observe that people describe the target more often in the false-incongruent condition (when the robot looks at the target) than in the false-congruent or false-no gaze conditions.

Results

Results described below are plotted in Figure 6. We observe a main effect of the factor 'described object' ($F(1,35) = 29.31, p < 0.001$) suggesting that people generally prefer to correct a sentence with respect to the competitor which has been identified by colour. Moreover, we find an interaction effect of the two factors 'described object' and 'gaze condition' ($F(2,70) = 16.04, p < 0.001$) indicating that gaze does indeed has an impact on the choice of referents. More precisely, we observe that people correct an utterance with respect to the target significantly more often when the robot looks towards the target (false-incongruent) than when it looks at the competitor or nowhere at all.

Another fact indicating that gaze is affecting reference resolution becomes apparent when analysing corrections in response to *true* robot utterances. Although we did not expect participants to correct true statements, we observed that in 15% of true-incongruent trials people corrected the robot with a sentence about the competitor. This suggests that people believed that the robot was indeed talking about the competitor that it looked at even though both the comparative and the colour referred to the target object.

The results from the correction analysis supports our hypothesis that people believe robot gaze to reflect its intention to talk about an object it looks at (Prediction 2). The inspection patterns in this experiment, though not reported, are very similar to the ones observed in Experiment 1 and strengthen the hypothesis that people integrate visual and linguistic references online, similar to human-human interaction.

[3] We are aware that ANOVA is not the optimal test for originally categorical data but to our knowledge this remains the standard analysis in most production studies.

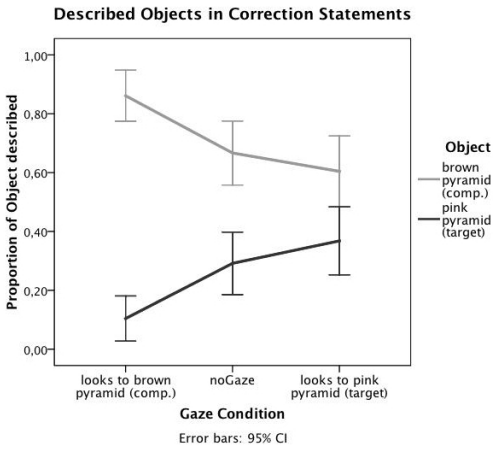


Figure 6: Proportion of objects described in response to false utterances, e.g. "The cylinder is bigger than the pyramid that is brown".

Conclusions

Neither task in the presented experiments required participants to exploit robot gaze, they rather required them to listen to the utterance while observing the object arrangements. Nonetheless, we see clear on-line evidence of gaze-following, despite a relatively high proportion of incongruent trials which might lead subjects to lose confidence in the robot's performance. We further observed that robot gaze as a visual cue can even override some linguistic cues. These results suggest that robot gaze even in this minimal form, being merely simulated by a moving stereo camera head, provides a visual cue that people respond to in the same automatic way that they respond to human gaze (Prediction 1).

We further present support for the hypothesis that cognitively plausible robot gaze is beneficial for comprehension (and thus for communication) by showing that people faster respond to congruent robot gaze and speech behaviour. To distinguish the purely visual component of robot gaze from its potentially referential meaning, we changed the task from a response time task in Experiment 1 to a production task in Experiment 2, in which people had to verbally correct the robot's statement. With no time pressure on people's responses, the shift of visual attention cannot be the crucial factor anymore. Instead, the produced correction statements reveal which referent listeners thought was 'intended' by the robot and confirm that this was influenced by robot gaze (Prediction 2).

The presented findings suggest that people integrate *visual reference* information derived from the robot's gaze with their on-line interpretation of robot speech. This finding is broadly consistent with Knoeferle and Crocker's (2007) "Coordinated Interplay Account" (CIA) of situated comprehension, but requires an extension to that model in which not only speech, but also speaker gaze, is used to both direct attention and visually ground utterance meaning during comprehension. Interestingly, just as Knoeferle and Crocker show that scene events can have priority over linguistic expectations, we similarly find evidence that speaker (robot) gaze can override linguistic cues about intended referents. This highlights the general importance of visual information (both the scene and speaker gaze) during situated language processing, and supports our more specific claim that cognitive models of real-time language-mediated gaze benefit situated communication with artificial agents such as robots.

Acknowledgments

The research reported of in this paper was supported by IRTG 715 "Language Technology and Cognitive Systems" funded by the German Research Foundation (DFG).

References

Altmann, G., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. In J. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (p. 347-386). NY: Psychology Press.

Beun, R., & Cremers, A. (1998). Object Reference in a Shared Domain of Conversation. *Pragmatics and Cognition*, *6*, 111-142.

Chris Moore, P. D., Philip J. Dunham (Ed.). (1995). *Joint Attention Its Origins and Role in Development*. LEA.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, *82*, B1-B14.

Hanna, J., & Brennan, S. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *JML*, *57*, 596-615.

Kiesler, S., Powers, A., Fussell, S., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robotlike agent. *Social Cognition*, *26*, 169-181.

Knoeferle, P., & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: evidence from eye-movements. *JML (Special issue: Language-Vision Interaction)*, *57*, 519-543.

Meyer, A., Sleiderink, A., & Levelt, W. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, *66*, B25-B33.

Schroeder, M., & Trouvain, J. (2001). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. In *4th isca workshop on speech synthesis.* Blair Atholl, Scotland.

Staudte, M., & Crocker, M. W. (2009). Visual Attention in Spoken Human-Robot Interaction. In *Proc. of the 4th ACM/IEEE Conference on HRI.* San Diego, USA.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632-1634.

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