Producing and resolving multi-modal referring expressions in human-robot interaction

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

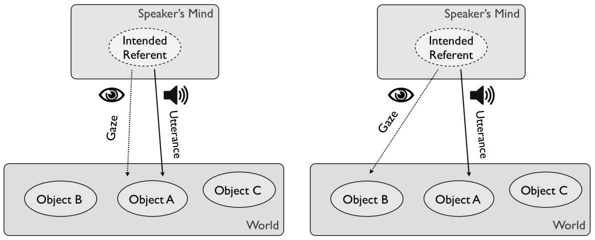
Keywords: human-robot interaction, grounding, multi-modal, reference resolution

Introduction

People exploit various information modalities to produce and resolve referring expressions (REs) in situated communication. While the given (visual) situation provides a context such that linguistic references can be grounded in the actual world, speakers and listeners themselves provide multi-modal information to further constrain and facilitate interpretation within this context. In addition to linguistic REs, cues such as pointing gestures and eye gaze constitute (possibly redundant or complementary) *visual references*. Ideally, all these cues identify the same object in the scene. However, it may occur that a person looks at a mug while saying ”Pass me the glass, please.” and the listener may be in doubt about what to do. Visual references thus have considerable impact on the resolution of linguistic REs and our aim is to investigate this relationship in more detail, in the context of situated human-robot communication.

While pointing gestures are often used deliberately, frequently substituting parts of a linguistic RE (when definite noun phrases, for instance, are replaced by deictic pronouns and an according pointing gesture (Bangerter, 2004)), gaze is mostly an uncontrolled cue that automatically accompanies our utterances. This view is supported by psycholinguistic studies that have shown, for instance, that referential speech and gaze are temporally extremely closely aligned. On one hand, *speakers* often look at what they intend 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). On the other 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.

Because of this close temporal coupling of gaze and speech, gaze cues can contribute to the resolution of linguistically underspecified REs. Hanna & Brennan (2007), for instance, have shown that listeners use speakers' gaze to identify a referent in the scene before the utterance unambiguously identifies that referent. Their study also showed that the speaker's gaze helps to identify possible referents even when the gaze is initially misleading (induced by the experimental setup). Subjects were able to establish a mapping of the speaker's gaze to their own visual scene and, thus, make use of the speaker's gaze early during comprehension. Among other studies, the results from Hanna & Brennan suggest that people infer intended referents from the speaker's gaze even though the initial physical response to that gaze is likely to be mostly reflexive visuospatial orienting (Friesen & Kingstone, 1998; Driver et al., 1999; Langton & Bruce, 1999).



(a) Congruent multi-modal reference to one object

(b) Incongruent multi-modal reference to two different objects

Figure 1: From intended referents to objects in the scene.

The above findings show that people produce and comprehend gaze cues on-line and that those are often interpreted as visual references augmenting linguistic REs (as shown in Figure 1a). In Staudte & Crocker (2009b), evidence was reported from two eye-tracking studies suggesting that people similarly follow robot gaze cues when listening to the robot's statements about a shared visual environment. The produced gaze cues were temporally aligned with the robot utterance according to the described temporal pattern of human RE production. We manipulated this gaze and speech alignment such that both modalities were not always congruent, i.e., the robot looked at one object while mentioning another (see Figure 1b). These studies have replicated the effect found by Driver and colleagues who showed that people instantly and reflexively follow the direction of a depicted pair of human eyes. While we used dynamic and yet simple robot gaze represented by a moving stereo camera (see Figure 2), we also observed reliable gaze following behaviour of our participants. Furthermore, we asked participants to press a button according to the correctness of the robot's utterance as soon as they could. These response time data, that in fact are in line with Hanna and Brennan's work, suggest that following these gaze cues also has an influence on utterance comprehension: Participants' responses to incongruent behaviour were significantly slower than to congruent behaviour and slower than to bare utterances that were not accompanied by gaze - even though responses were regarding only the statement's validity.

We therefore hypothesize that people indeed integrate cognitively motivated robot gaze with the actual robot utterance when resolving REs. That is, both gaze cues and the utterance are combined in a single reference resolution mechanism. Supporting evidence was found in a follow-up experiment where it was shown that people not only (automatically) follow robot gaze to objects in a scene but that robot gaze also influences what people assume the intended referent is (Staudte & Crocker, 2009a). In this paper, we present the results of this experiment and discuss how our findings contribute to understanding how gaze as an automated visual reference mechanism interacts with linguistic references.

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 (temporarily ambiguous) statements about this scene[1] (Figure 2). These videos varied with respect to the robot's sentence validity (true/false) and the robot's gaze congruency (congruent/incongruent/absent). Our participants were eye-tracked while observing these videos. The task required participants to give a corrected sentence of a robot's utterance when they thought that the robot had made a mistake. Specifically, we observe participants in response to a false robot utterance that is accompanied by either incongruent, congruent or no robot gaze.

The reported results are two-fold: Firstly, the observed human gaze behaviour in response to the robot's visual and linguistic references replicates the findings of Staudte & Crocker (2009b) which employed a simple judgment task. Secondly, the produced correction statements provide insight about what people understood to be the robot's intended referent when its utterances were inaccurate and whether this was influenced by the robot's gaze.

Methods

Task & Procedure

Thirty-six native speakers of German, mainly students enrolled at Saarland University, took part in this study (12 males, 24 females). In this experiment, participants were told that Robbie, the robot, is an intelligent system that looks around and describes parts of the scene and that it may produce various mistakes in doing so. They were instructed to give a corrected sentence of the robot's utterance when they thought that the robot had made a mistake. Participants were further told that the overall goal of the experiment was to provide the robot with feedback for its learning mechanism so it could avoid making the same mistakes. 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. Recording of participants' utterances started with the trial onset and ended when participants pressed a button to continue.

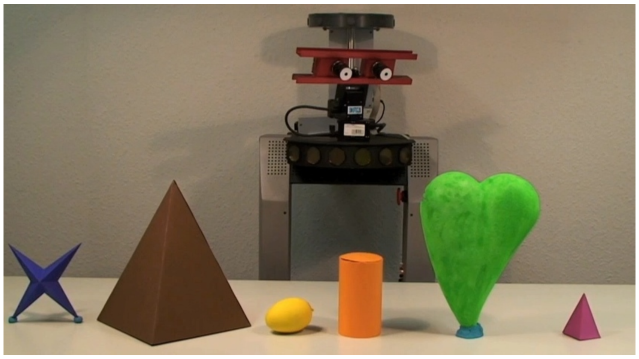


Figure 2: Sample scene with sample utterance: ”The cylinder is taller than the pyramid that is pink”

Materials

The robot utterances were originally in German and of the following form: ”The cylinder is taller than the pyramid that is pink.” While the robot utters such a sentence, it looks towards the cylinder (anchor) first, and then towards the pink pyramid (target). A set of 24 items was used so each participant saw four different items in each of the six conditions. 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 ”taller” to ”shorter” 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 (target vs competitor). 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, thus mimicking human production behaviour (see Figure 3 for approximate timing of the robot's behaviour)

Our scenes provide two potential referents for the final noun (e.g. two pyramids of different sizes and colours) one of which the robot then mentions explicitly by naming its colour. While the small pink pyramid (target) matches the example description of the scene (Figure 2), the big brown pyramid (competitor) does not. Note that the comparative is a cue pre-

[1] 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)

Table 1: Linguistic and visual references to objects in three congruency conditions for a false sentence, e.g. ”The cylinder is taller than the pyramid that is brown” where the small pink pyramid would be considered as target.

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

dicting the target object as actual referent, e.g., for the partial sentence ”The cylinder is taller than” the small pink pyramid is the target object since it is indeed shorter than the cylinder while the big brown pyramid is not. The mentioned colour of the object finally determines the factor *statement truth*: when the target is mentioned the statement is correct whereas mentioning the competitor results in an incorrect statement. The second factor is robot *gaze congruency*: Gaze is considered *congruent* with the utterance when both modalities refer to the same object in the scene and *incongruent* when gaze and utterance refer to different objects in the scene (Figure 1). The third congruency level is the absence of gaze such that only the utterance can convey a reference. The manipulation of both factors - statement validity and congruency - results in six conditions per item.

Because we want to mainly analyse the correction statements participants produce, false robot utterances are of particular interest in this experiment. As already mentioned, there are two cues in the robot utterance identifying the correct referent. The first cue is the comparative (*taller than* or *shorter 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 *taller* than the *brown* pyramid. Thus, people can repair this utterance by changing either the comparative or the colour adjective. The no-gaze-condition provides a baseline concerning the bias towards either repair in the absence of gaze. When robot gaze is present it increases the visual saliency of one of the potential referents: either it supports the mentioned object (identified by colour) or it supports the alternative object (identified by the comparative, not colour). Details on referential variation for the three *false* conditions are shown in Table 1.

Analysis

For the analysis of the corrections, we annotated the produced sentences with respect to which object was described (in response to false robot utterances only, i.e. considering only the conditions shown in Table 1). The three categories assigned to responses were *Target*, *Competitor* and *Else* (no correction given or described one or more different objects). Each response category is thus coded as a binary variable (e.g. the target has been described in the correction sentence or not). Since participants almost always either produced a sentence containing the target or the competitor, both response categories *Target* and *Competitor* are nearly complementary.

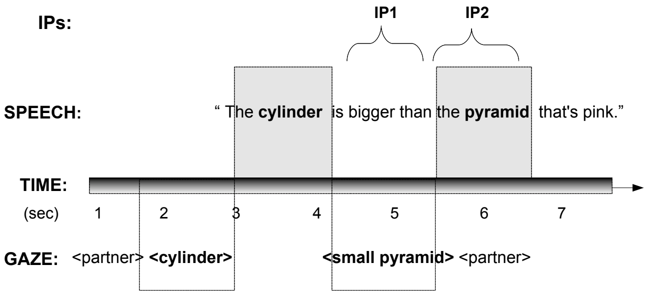


Figure 3: Approximate timing of robot behaviour. The two upper boxes mark the linguistic REs (nouns) while the lower boxes depict the timing of the visual references, i.e., robot gaze.

We also recorded people’s eye-movements during trials in order to compare participant behaviour in this study with the behaviour observed in previous studies. 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. We further segmented the video/speech stream into two Interest Periods (IP) as depicted in Figure 3. 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 (starting at when the camera reaches the target object) as well as verbal content preceding the target noun phrase (e.g. ”taller than”). IP2 stretches from the target noun onset to offset (mean duration of 674ms). Consecutive fixations within one IA were pooled as one inspection. We compute proportions of inspections per IA within an IP and 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. Since sentence truth does not play a role in IP1 and IP2 (because sentence truth cannot be determined until the occurrence of the sentence-final adjective), 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”.

Predictions

We hypothesize that the effect of robot gaze on people’s visual attention is due to the assumption that robot gaze is expressing some kind of intentionality and it consequently, similar to human gaze, elicits predictions about the intended referent of the speaker. If this is the case, we predict that robot gaze not only has an effect on how fast references are resolved but also on which object is believed to be the referent of the utterance. More precisely, people will then 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. If robot gaze, on the other hand, directs

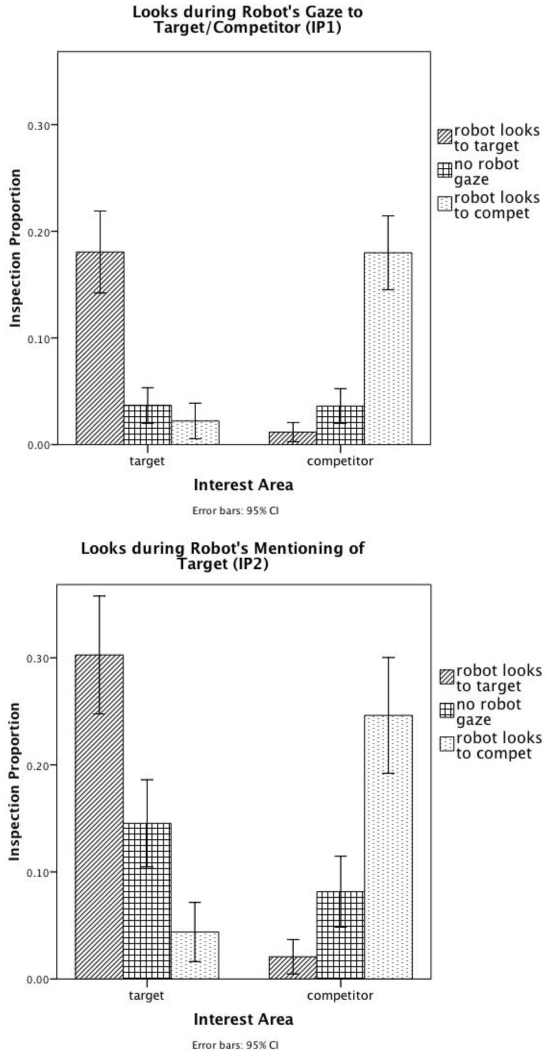


Figure 4: Inspections on target/competitor per condition and IP.

visual attention towards an object without contributing referential meaning, we expect that people's repair pattern in the gaze-conditions will not differ significantly from the no-gaze condition.

Results

Eye Movements

The findings on people's fixations during the experiment replicate the findings from our previous experiments. That is, people robustly follow the robot's gaze and speech to objects in the scene irrespective of the type of task they are given.

More precisely, in IP1, when the robot looks towards either the target or the competitor, we observe a main effect of robot gaze and a significant interaction between gaze direction and IA ($F_{gaze}(2,70) = 26.77, p < 0.001$, $F_{interaction}(2,70) = 66.39, p < 0.001$), i.e., people clearly follow this gaze and inspect the according IA. In IP2, when the robot mentions the target noun, the main effect of robot gaze and the interaction effect remain ($F_{gaze}(2,70) = 4.7, p < 0.05, p < 0.001$, $F_{interaction}(2,70) = 55.34, p < 0.001$). Moreover, we now find a main effect of IA. That is, people inspect the target object, which is coherent with the uttered sentence so far, generally more often than the competitor which is not coherent with the comparative. This tendency is particularly obvious in the no-gaze condition where a pairwise post-hoc comparison shows a significant difference between target and competitor. Interestingly, all three conditions within one IA are now pairwise significantly different. That is, while the robot makes an (partial/ambiguous) RE people look more at the target, for instance, when the robot had *previously* gazed at the target than when the robot showed no gaze behaviour at all. More importantly, people inspect a potential referent (e.g. the target) less when the robot gazed at a different object (competitor) then when there is no robot gaze.

Table 2: Logistic regression model for response category *Target* with separate subject and item analyses: glm(formula = cbind(target, competitor) ∼ GazeCondition, family = ”binomial”, data = bySubject).

Coefficients (bySubj)	Estimate	SE	z value	$Pr(>\|z\|)$
(Intercept) (fc)	-2.112	0.273	-7.727	<0.001
fi	1.617	0.324	4.987	<0.001
fn	1.286	0.330	3.895	<0.001
(byItem)	Estimate	SE	z value	$Pr(>\|z\|)$
(Intercept) (fc)	-2.112	0.273	-7.727	<0.001
fi	1.617	0.324	4.987	<0.001
fn	1.286	0.330	3.895	<0.001

Table 3: Logistic regression model for response category *Competitor* with separate subject and item analyses: glm(formula = cbind(competitor, target) ∼ GazeCondition, family = ”binomial”, data = bySubject).

Coefficients (bySubj)	Estimate	SE	z value	$Pr(>\|z\|)$
(Intercept) (fc)	2.112	0.273	7.727	<0.001
fi	-1.617	0.324	-4.987	<0.001
fn	-1.286	0.330	-3.895	<0.001
(byItem)	Estimate	SE	z value	$Pr(>\|z\|)$
(Intercept) (fc)	2.112	0.273	7.727	<0.001
fi	-1.617	0.324	-4.987	<0.001
fn	-1.286	0.330	-3.895	<0.001

Sentence Production

The response category *Else* was found in 3.47% of the *false*-trials and was treated as missing values in the analyses described below. The mean proportion of corrections involving the target/competitor are depicted for each condition in Figure 5. While the shown proportions are given for visualisation purposes, we did not use those for the analysis. Instead, we fitted a logistic regression model to out data (for response categories target and competitor separately) with one categorical

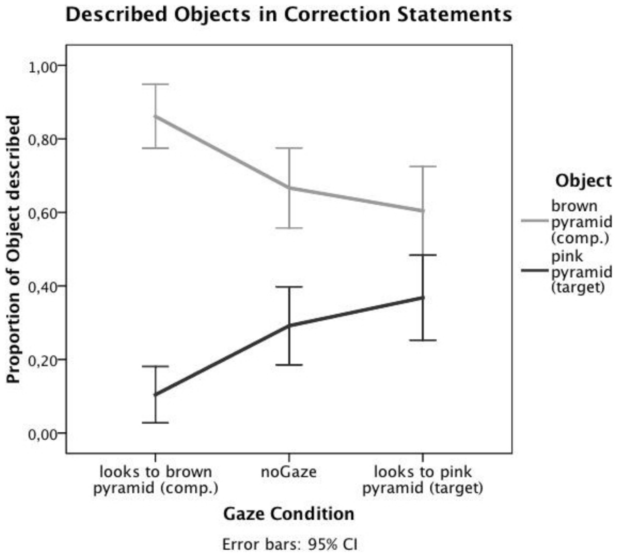


Figure 5: Proportion of objects described in response to false utterances, e.g. ”The cylinder is taller than the pyramid that is brown”, where the colour adjective identifies the competitor.

predictor (’gaze congruency’ with the levels *false-congruent (fc)*, *false-no gaze (fn)* and *false-incongruent (fi)*. The results given in Tables 2 and 3 for subject and item analyses respectively[2] show a main effect of gaze-congruency (all models yielded p-values below 0.001). Specifically, we observed that people corrected an utterance using the target (i.e. change the colour) significantly less often when the robot looked towards the competitor (false-congruent) or nowhere than when it actually looked at the target. These results are depicted by the black line in Figure 5. In contrast, people mentioned the competitor (i.e. changed the comparative) in their correction statements even more often when the robot also looked at the competitor than when it looked at the target or nowhere at all, as is depicted by the grey line.

A general preference to build a corrected sentence about the competitor (which has been linguistically identified by the mentioned colour) is clearly visible in the no-gaze baseline condition (central condition in Figure 5). In the absence of robot gaze, people used the competitor - and changed the comparative - in almost 67% of their correction statements. This general preference for the more explicitly mentioned object (colour match) remained dominant in all three gaze conditions due to two possible reasons. Firstly, it has been shown that people prefer to use absolute (shape and colour) to relative features (size, location) for the production of REs (Beun & Cremers, 1998). And secondly, gaze is frequently incongruent in our stimuli (and considered incorrect) whereas speech is always fluent and clear which may induce a general competence bias towards explicitly mentioned objects.

Another fact indicating that gaze affects 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 mentioned comparative and mentioned colour identified the target object.

Discussion & Conclusions

We observed clear on-line evidence of gaze-following across our experiment series, despite a relatively high proportion of incongruent trials that could have led subjects to lose confidence in the robot’s performance. The fixation results show that robot gaze has an even stronger influence on people’s visual attention than other linguistic cues that typically elicit fixations to potential referents during incremental reference resolution (see Staudte & Crocker (2009a) for details). 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 initially respond to in a similarly automatic way that they respond to human gaze.

Previous findings from the psychological literature have suggested that this response behaviour is not unique to human (or robot) gaze but that other attention directing cues such as arrows trigger similar reflexive behaviour. However, Ristic etal. (2007) have shown that a gaze cue primes a location more reliably than arrows where the priming effect is subject to colour congruency between the arrow and the actual target stimulus. According to the authors, this indicates that the attention effect for gaze is more strongly reflexive as for arrows. An additional or alternative explanation for this reliable attention effect of eyes/gaze may be related to intentional gaze processing (Castiello, 2003; Bayliss, Paul, Cannon, & Tipper, 2006; Becchio, Bertone, & Castiello, 2008). Bayliss and colleagues (2006) have shown, for instance, that a visual referent that was looked at by another person receives higher likability scores than a not-looked at object. Another series of studies conducted by Castiello (2003) has shown, for instance, that people even infer motor intentions from an actor’s gaze. Based mainly on these results, Becchio and colleagues argue that gaze potentially enriches the representation of a visual referent and they propose a ”mechanism that allows transferring to an object the intentionality of the person who is looking at it” which they call ”intentional imposition”.

The production results of our study also suggest that robot gaze not only triggers reflexive visuospatial orienting but that people use robot gaze to infer the *intended* referents. Since in the presented study participants were asked to verbally correct the robot’s statement in a self-paced setting with no time pressure on their responses, the reflexive shift of visual attention alone cannot account for the chosen object that people describe in their corrected sentences, indicating which object they understood to be the intended referent. It is also worth mentioning, that the reflexive attention shift effect as

[2] We obtained similar results fitting a logistic mixed-effect model but refrained from using the resulting p-values as there is no MCMC-method (yet) for validating p-values for binomial data.

discussed above occurs only if the target stimulus appears within a very short time window after the cue (e.g. 100ms in Langton & Bruce (1999)). Our eye-movement data, however, clearly show that people are still influenced in IP2 by the previously performed robot gaze cue even though it is uninformative (0.5 probability for predicting the linguistic referent) which indicates an effect of robot gaze that is neither purely reflexive nor related to strategic behaviour.

Our data therefore provide further support for the view that gaze is indeed processed as an intentional cue as suggested by Becchio etal. Moreover, our results suggest that intentional gaze processing is applied not only to human eyes but also when faced with an extremely simple realisation of robot gaze (represented by a moving stereo camera).

We conclude that we have successfully shown that robot gaze which is aligned to the robot's utterance in a cognitively-derived manner, augments the uttered RE and influences the resolution thereof. We further conclude that the presented experimental design allows us to manipulate the production of multi-modal REs in situated communication while being able to investigate how people process these REs.

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 Pirita Pyykkonen for valuable feedback.

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.

Bangerter, A. (2004). Using Pointing and Describing to Achieve Joint Focus of Attention in Dialogue. *Psychological Science*, *15*, 415-419.

Bayliss, A., Paul, M., Cannon, P., & Tipper, S. (2006). Gaze cuing and affective judgments of objects: I like what you look at. *Psychonomic Bulletin & Review*, *13*, 1061-1066.

Becchio, C., Bertone, C., & Castiello, U. (2008). How the gaze of others influences object processing. *Trends in Cognitive Science*, *12*, 254-258.

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

Castiello, U. (2003). Understanding Other People's Actions: Intention and Attention. *Journal of Experimental Psychology*, *29*, 416-430.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze Perception Triggers Reflexive Visuospatial Orienting. *Visual Cognition*, *6*, 509-540.

Friesen, C., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, *5*, 490-495.

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.

Langton, S. R., & Bruce, V. (1999). Reflexive Visual Orienting in Response to the Social Attention of Others. *Visual Cognition*, *6*, 541-567.

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

Ristic, J., Wright, A., & Kingstone, A. (2007). Attentional control and reflexive orienting to gaze and arrow cues. *Psychonomic Bulletin & Review*, *14*, 964-969.

Staudte, M., & Crocker, M. W. (2009a). The effect of robot gaze on processing robot utterances. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society.* Amsterdam, Netherlands.

Staudte, M., & Crocker, M. W. (2009b). Visual Attention in Spoken Human-Robot Interaction. In *Proceedings 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.*