# Joint Attention
# in
# Spoken Human-Robot Interaction

Dissertation
zur Erlangung des akademischen Grades eines
Doktors der Philosophie
der Philosophischen Fakultäten
der Universität des Saarlandes

vorlegt von

Maria Staudte

aus Berlin

Saarbrücken, 2010

| | |
|---|---|
| Dekan: | Prof. Dr. Erich Steiner |
| Berichterstatter: | Prof. Dr. Matthew W. Crocker |
| | Prof. Dr. Gerry T. M. Altmann |
| Tag der letzten Prüfungsleistung: | 12. Juli 2010 |

# Abstract

Gaze during situated language production and comprehension is tightly coupled with the unfolding speech stream – speakers look at entities before mentioning them (Griffin, 2001; Meyer et al., 1998), while listeners look at objects as they are mentioned (Tanenhaus et al., 1995). Thus, a speaker's gaze to mentioned objects in a shared environment provides the listener with a cue to the speaker's focus of visual attention and potentially to an intended referent. The coordination of interlocutor's visual attention, in order to learn about the partner's goals and intentions, has been called *joint attention* (Moore and Dunham, 1995; Emery, 2000). By revealing the speakers communicative intentions, such attentional cues thus complement spoken language, facilitating grounding and sometimes disambiguating references (Hanna and Brennan, 2007).

Previous research has shown that people readily attribute intentional states to nonhumans as well, like animals, computers, or robots (Nass and Moon, 2000). Assuming that people indeed ascribe intentional states to a robot, joint attention may be a relevant component of human-robot interaction as well. It was the objective of this thesis to investigate the hypothesis that people jointly attend to objects looked at by a speaking robot and that human listeners use this visual information to infer the robot's communicative intentions.

Five eye-tracking experiments in a spoken human-robot interaction setting were conducted and provide supporting evidence for this hypothesis. In these experiments, participants' eye movements and responses were recorded while they viewed videos of a robot that described and looked at objects in a scene. The congruency and alignment of robot gaze and the spoken references were manipulated in order to establish the relevance of such gaze cues for utterance comprehension in participants.

Results suggest that people follow robot gaze to objects and infer referential intentions from it, causing both facilitation and disruption of reference resolution, depending on the match or mismatch between inferred intentions and the actual utterance. Specifically, we have shown in Experiments 1-3 that people assign attentional and intentional states to a robot, interpreting its gaze as cue to intended referents. This interpretation determined how people grounded spoken references in the scene, thus, influencing overall utterance comprehension as well as the production of verbal corrections in response to false robot utterances. In Experiments 4 and 5, we further manipulated temporal synchronization and linear alignment of robot gaze and speech and found that substantial temporal shifts of gaze relative to speech did not affect utterance comprehension while the order of visual and spoken referential cues did. These results show that people interpret gaze cues in the order they occur in and expect the retrieved referential intentions to be realized accordingly. Thus, our findings converge to the result that people establish joint attention with a robot.

# Zusammenfassung

Die Blickrichtung des Menschen ist eng mit Sprachproduktion und Sprachverstehen verknüpft: So schaut ein Sprecher in der Regel auf ein Objekt kurz bevor er es nennt, während der Blick des Hörers sich beim Verstehen des Objektnamens darauf richtet (Griffin, 2001; Meyer et al., 1998; Tanenhaus et al., 1995). Die Blickrichtung des Sprechers gibt dem Hörer also Aufschluss darüber, wohin die Aufmerksamkeit des Sprechers gerade gerichtet ist und worüber möglicherweise als nächstes gesprochen wird. Wenn jemand dem Blick seines Gegenübers folgt, um herauszufinden was dieser für Ziele oder Absichten hat, spricht man von gemeinsamer Aufmerksamkeit (Joint Attention, bzw. Shared Attention, wenn beide Gesprächspartner ihre Aufmerksamkeit bewusst koordinieren, Moore and Dunham, 1995; Emery, 2000). Der Blickrichtung des Sprechers zu folgen kann demnach nützlich sein, da sie häufig seine Absichten verrät. Sie kann sogar das Sprachverstehen erleichtern, indem zum Beispiel referenzierende Ausdrücke mit Hilfe solcher visuellen Informationen disambiguiert werden (Hanna and Brennan, 2007).

Darüber hinaus wurde in der Vergangenheit gezeigt, dass Menschen häufig nicht nur Menschen, sondern auch Tieren und Maschinen, wie zum Bespiel Robotern, Absichten oder Charakterzüge zuschreiben (Nass and Moon, 2000). Wenn Robotern tatsächlich die eigentlich menschliche Fähigkeit, Ziele oder Absichten zu haben, zugeordnet wird, dann ist davon auszugehen, dass gemeinsame Aufmerksamkeit auch einen wichtigen Bestandteil der Kommunikation zwischen Mensch und Roboter darstellt. Ziel dieser Dissertation war es, die Hypothese zu untersuchen, dass Menschen versuchen Aufmerksamkeit mit Robotern zu teilen, um zu erkennen was ein Roboter beabsichtigt zu sagen oder zu tun.

Wir stellen insgesamt fünf Experimente vor, die diese Hypothese unterstützen. In diesen Experimenten wurden die Augenbewegungen und Antworten, beziehungsweise Reaktionszeiten, von Versuchspersonen aufgezeichnet, während letztere sich Videos anschauten. Die Videos zeigten einen Roboter, welcher eine Anordnung von Objekten beschrieb, während er seine Kamera auf das ein oder andere Objekt richtete um Blickrichtung zu simulieren. Manipuliert wurde die Kongruenz der Verweise auf Objekte durch Blickrichtung und Objektnamen, sowie die Abfolge solcher Verweise. Folglich konnten der Informationsgehalt und die relative Gewichtung von Blickrichtung für das Sprachverstehen bestimmt werden.

Unsere Ergebnisse belegen, dass Menschen tatsächlich dem Roboterblick folgen und ihn ähnlich interpretieren wie die Blickrichtung anderer Menschen, d.h. Versuchspersonen leiteten aus der Blickrichtung des Roboters ab, was dessen vermeintliche (sprachliche) Absichten waren.

Insbesondere zeigen die Experimente 1-3, dass Versuchspersonen die Blickrichtung des Roboters als Hinweis auf nachfolgende referenzierende Ausdrücke verstehen und dementsprechend die Äußerung des Roboter speziell auf jene angeschauten Objekte beziehen. Dies führt zu verkürzten Reaktionszeiten wenn die Verweise auf Objekte durch Blickrichtung und Objektnamen übereinstimmen, während widersprüchliche Verweise zu verlängerten Reaktionszeiten führen. Dass Roboterblick als Ausdruck einer (sprachlichen) Absicht interpretiert wird, zeigt sich auch in den Antworten, mit denen Versuchspersonen falsche Aussagen des Roboters korrigierten. In den Experimenten 4-5 wurde außerdem die Anordnung der Verweise durch Blick und Sprache manipuliert. Während die genaue zeitliche Abstimmung der Verweise den Einfluss von Roboterblick nicht mindert, so scheint die Reihenfolge der Verweise entscheidend zu sein. Unsere Ergebnisse deuten darauf hin, dass Menschen Absichten aus den Verweisen durch Blickrichtung ableiten und erwarten, dass diese Absichten in derselben Anordnung umgesetzt werden. Insgesamt lassen unsere Ergebnisse also darauf schließen, dass Menschen versuchen, ihre Aufmerksamkeit gemeinsam mit Robotern zu koordinieren, um das Sprachverstehen zu erleichtern.

# Acknowledgments

During the exciting and stimulating but sometimes also long and exhausting process of producing a dissertation, some people have made substantial contributions to this work and deserve my deepest gratitude.

First and foremost, I want to say that I am extremely thankful to my supervisor Matthew Crocker without whom I could not have written this thesis. If it had not been for his interest in my somewhat unconventional ideas almost three years ago, I certainly would not have become what I am now. While giving me the freedom to do what I liked, he has always been very supportive, patiently discussing upcoming ideas and problems.

I am also grateful to Geert-Jan Kruijff and Hans Uszkoreit for making me become interested in human-robot interaction in the first place and for supplying me with hardware and advice, especially in the initial phase.

Many thanks also go to my colleagues and friends Judith Köhne, Emilia Ellsiepen, Afra Alishahi, Pirita Pyykkönen, Mark Buckley, Berry Claus, Juliane Steinberg und Helene Kreysa for making the past three and a half years a pleasant time and for always encouraging and supporting me in one way or another. I am also grateful for their professional advice and the help in proof-reading and, finally, also for providing me with food and the like in the final phase of thesis writing.

I also would like to thank the IRTG 715 ”Language Technology and Cognitive Systems” for providing a fruitful and friendly environment for graduating as well as for the financial support throughout this time. Of course, I could not have managed all the administrative tasks involved in experimenting, traveling, and celebrating without Claudia Verburg who has always been extremely helpful.

Thanks to my family, especially my parents, who always supported me in any possible way and who seem to have endless faith in me. Vielen Dank für Euer Vertrauen, Eure Liebe und dass Ihr immer hinter mir steht!

Finally, I would like to express my deepest gratitude to my love Michael. Thank you so much for loving and always supporting me!

# Contents

# 1. Introduction

According to an old and widespread proverb, the eyes are windows to the soul. The validity of this statement has, at least to some extent, been supported by a large body of previous psychological and psycholinguistic research. Baron-Cohen et al. (1997a) state, for instance, that

> "we showed that a small number of other mental states can also be read from direction of gaze. These include desire, refer, and goal (Baron-Cohen, Campbell, Karmiloff-Smith, Grant, & Walker, 1995). That is, our natural reading of gaze directed at a specific object is in terms of a person's volitional states. This should come as no surprise, since we tend to look at what we want, and to what we are referring, and at what we are about to act upon." (p.312)

The primary function of directing gaze is certainly related to the act of *seeing*. To fixate something or somebody lets us inspect it or her in greater detail. Additionally, gaze in communication reflects numerous different processes and responds to many cues. It conveys, for instance, information about emotions, goals or desires: The eyes of the partner may express a certain emotion and where they are directed during talking may express a certain attitude (Argyle and Dean, 1965; Dovidio and Ellyson, 1982; Baron-Cohen et al., 1997b). Depending on the occurrence and duration, direct eye contact (also called *mutual gaze*) with a partner may appear threatening or dominant, while averted gaze may appear submissive or arrogant (Dovidio and Ellyson, 1982). Moreover, gaze may help organizing communication. Mutual gaze, for example, can be a useful cue for a listener to signal that she will take a speaking turn (Kendon, 1967). In addition to these meta-linguistic functions of gaze, it can also reflect information that is directly linked to the content of a spoken utterance. A deictic expression accompanied by a glance towards a certain object may be a valid and comprehensible reference for a listener in face-to-face communication (Clark and Krych, 2004). Thus, a listener seems to be able to link the spoken reference to the object which is in focus of the speaker's visual attention.

The emphasis of this thesis is precisely on this linguistically relevant role of gaze, potentially communicating attentional states and referential intentions which may influence both production as well as comprehension of an utterance.

Previously, gaze has been widely studied as an indicator for overt visual attention during language processing and it was shown that where we look is closely related to what we say and understand. Studies have revealed, for instance, that speakers look at entities roughly 800msec - 1sec. before mentioning them (Griffin, 2001; Meyer et al., 1998), while listeners inspect objects as soon as 200-400msec after the onset of the corresponding referential noun (Tanenhaus et al., 1995; Allopenna et al., 1998). This shows that eye gaze during situated language production and comprehension is tightly coupled with the unfolding speech stream. In face-to-face communication, the speaker's gaze to mentioned objects in a shared environment also provides the listener with a visual cue as to the speaker's focus of (visual) attention (Flom et al., 2007). Following this cue in order to attend to the same object as the partner has been dubbed *joint attention* (by Emery, 2000, and others as reviewed in Section 2.4). By revealing a speaker's focus of visual attention, such gaze cues potentially offer the listener valuable information to ground and sometimes disambiguate referring expressions, to hypothesize about the speaker's communicative intentions and goals and, thus, to facilitate comprehension (e.g., Clark and Krych, 2004; Hanna and Brennan, 2007).

It is an interesting question whether such gaze behavior is unique to human-human interaction – possibly hinging on common biological and cognitive mechanisms – or whether such gaze cues play a similar role in human-machine interaction. Previous research has shown that people readily attribute intentional states and personality traits to non-humans as well, like animals or artificial agents such as robots (see e.g. Nass and Moon, 2000, or Kiesler et al., 2008, for overviews). Assuming that people indeed ascribe intentional states (or at least goal-directedness) to robots, joint attention may be an important component of human-robot interaction as well.

Before addressing this issue, we will briefly review the most relevant findings on gaze and its coupling to language as well the role of gaze for joint attention. We then explain to what extent the insights on human gaze have been used to enrich human-computer interaction and which important questions remain to be investigated. Finally, we discuss whether robot gaze can in principle fulfill similar functions as human gaze and how we have examined this issue, before giving an overview of the theoretical and experimental work presented in this thesis.

## 1.1. Use of Gaze during Language Processing

Since language is often vague and ambiguous, additional non-verbal cues supporting and augmenting the conveyed message or the retrieval of information are potentially useful in face-to-face communication. While cues like pointing generally complement spoken language and are potentially useful to ground and disambiguate an utterance in the scene (Hanna and Brennan, 2007), gaze seems to be a special one among such non-verbal cues: Gaze is permanently available since people constantly use and move their eyes even when their gaze is not related to language production or comprehension. Further, gaze is extremely diverse in its expressiveness conveying various emotions and other mental states as suggested by a large body of research (see, e.g., Adams and Kleck, 2003; Baron-Cohen et al., 1997b; Dovidio and Ellyson, 1982).

The close coupling of language and gaze has been established in a number of studies (Tanenhaus et al., 1995; Allopenna et al., 1998; Meyer et al., 1998; Altmann and Kamide, 1999; Griffin and Bock, 2000; Altmann and Kamide, 2004; Knoeferle et al., 2005). On one hand, where people look is driven by what they hear or say (linguistic processing), and on the other hand, it is driven by what they see (visual processing, Henderson, 2003) which includes speakers' gaze as a visual cue. Possibly because of this systematic and automatic coupling listeners can interpret speakers' eye-movements on-line as visual references.

Whether, and how, the close alignment of visuo-linguistic processes helps listeners to comprehend utterance content, is subject to ongoing research (Crocker et al., in press). Previous studies on joint attention suggest that people do indeed monitor and use each others gaze and speech in face-to-face communication to rapidly ground and resolve spoken utterances with respect to a common environment (Moore and Dunham, 1995; Clark and Krych, 2004; Tomasello and Carpenter, 2007). That is, where a speaker looks may constraint the domain of interpretation for the listener (Hanna and Brennan, 2007) and where a listener looks may tell the speaker that she has misunderstood such that the speaker may decide to repeat or to further specify a referring expression (Clark and Krych, 2004).

Thus, referential gaze is closely aligned to speech – and the question arises what happens when this alignment is disrupted. That is, how do people deal with gaze cues that are incongruent or miss-aligned with the spoken utterance? Such situations occur, for instance, when misunderstandings lead to the use of inappropriate objects names in both human-human or human-computer interaction. In the latter, incongruent multi-modal references (i.e., combined linguistic and visual cues) could easily be caused by an agent's "mis-programmed" gaze movements or errors in its object recognition. Insights on how inappropriate co-occurrences of gaze and speech cues are resolved offer the potential to illuminate the nature of gaze influence as well as the integration process of information provided through different modalities such as language and vision.

## 1.2. Mechanisms behind the Use of Gaze as a Cue

Listeners may use speaker gaze as a timely cue to utterance content, possibly because of the tight coupling of gaze and speech mentioned above. In order to understand why and how people use each others gaze as referential cues, the notion of visual attention is essential. It helps to establish and understand the connection between eye gaze and its referents in the external world. Allocation of visual attention allows more detailed inspection of one aspect in the environment (*selectivity*) while limiting processing of other (visual) information (*capacity limitation*) (Bundesen, 1990; Desimone and Duncan, 1995). That is, an entity that is being looked at is typically in the focus of visual attention, allowing investigation of the entity's visual features in greater detail. However, visual focus and visual attention can be dissociated such that a person may direct her visual focus (gaze, also called overt attention) towards an object while she already shifts (covert) attention to another entity (Posner, 1980). While covert visual attention can be shifted without shifting eye-gaze, the opposite is not necessarily the case (Hoffman and Subramaniam, 1995; Posner, 1980). That is, gaze shifts are preceded by covert visual attention shifts. Consequently, following the interlocutors' overt gaze shifts typically reveals information about what she is or has been visually attending to and may result in joint attention to the entity in focus.

Following Emery (2000), we consider *joint attention* to occur when a subject follows another subject's gaze to mutually attend to an entity, while possibly inferring her referential intentions. Joint attention presupposes that the gazer has attentional states such that the follower has reason to consider the looked-at entity as relevant. Further, the term *shared attention* is used to refer to a similar phenomenon which additionally involves intention sharing: One person *intentionally* directs another person's gaze to an object by looking at this object, in order to communicate goals for cooperating in task completion or just to share the experience (Emery, 2000).

It was previously investigated what kind of attention shift gaze cues may elicit (potentially resulting in joint attention) and to which extent eye gaze influences the assignment of attentional and intentional states to the gazer. Results suggest that people follow gaze, and infer mental states from it, since they learned that other human beings are similar to themselves and that *seeing* something with one's eyes means attending to it (Baron-Cohen et al., 1995; Meltzoff and Brooks, 2007). This attribution of perceptual (seeing) mental states as well as volitional mental states (desires, goals) to oneself and to others is a prerequisite for building a *theory of mind* (Baron-Cohen, 1995). To have a theory of mind means to use knowledge about mental states in general, and about epistemic mental states (believing, knowing, pretending) in particular, in a "theory-like" way to reason about and predict actions of others (Baron-Cohen, 1995, p.51ff). Therefore, having a theory of mind of others implies the capacity to interpret other's behavior in terms of mental states (Premack and Woodruff, 1978; Frith and Frith, 2005). In other words, an individual can draw inferences about why another person behaves likes she does because she can imagine what goals and intentions have elicited this behavior. Thus, it seems that the development of a theory of mind is a crucial component underling the use of gaze as a cue to (referential) intentions. It follows that the role of gaze in language production and comprehension is similarly closely related to our understanding of the partner as an intentional being since the interpretation of gaze as a cue to intended referents requires the assignment of perceptual and volitional states to the gazer.

## 1.3. Robot Gaze in Interaction with a Person

Despite the generally growing interest in human-computer/human-robot interaction (HCI/HRI) to incorporate natural gaze mechanisms, the psycholinguistic findings concerning referential gaze described above have not been systematically investigated. Rather, previous work on gaze in in HCI/HRI has concentrated largely on the general appearance of the agent and what competences and characteristics people intuitively ascribe to agents featuring certain gaze behaviors. Kanda and colleagues (2001), for instance, equipped their robot with very basic gaze movements and observed that people generally found the interaction more enjoyable than when the robot showed no gaze movements. Thus, robot gaze can, on one hand, improve agreeableness of HRI. On the other hand, robot gaze can be dysfunctional and disturb smooth interaction. In the same study, Kanda et al. (2001) found that the robot's crude gaze movements resulted in a lower performance judgement revealed by a post-experiment questionnaire. Similarly, Sidner and colleagues (2005) found that participants judged the robot they had to interact with to be less 'reliable' when it showed gaze (or head) movement. It was also found, however, that participants became more non-verbally engaged in the conversation with a robot when it showed gaze behavior (see also Wang et al., 2006). That is, participants produced more head nods and gaze cues in response to a robot that also produced such cues.

Another study conducted by Cassell et al. (1999c) revealed that the usage of mutual gaze is a function of turn-coordination and discourse information structure. This finding was partially used to implement and test a model for gaze production on a virtual agent (Cassell et al., 1999a; Cassell and Thórisson, 1999) and a robot (Mutlu et al., 2006). Results showed that such gaze behavior elicited positive impressions (agent was perceived as helpful and lifelike) and improved people's ability to later recall facts mentioned in this interaction.

The above mentioned studies on HCI suggest that gaze in one way or the other affects the impression a person or agent makes. Since appropriate and inappropriate robot behavior positively and negatively influences HCI, respectively, improvement of agent gaze behavior requires more information on human gaze production and processing. Psycholinguistic evidence reported in Sections 1.1 and 1.2, for instance, show that gaze can provide additional information that helps to quickly link the accompanying utterance to the world and guide attention accordingly. There has been limited research in HCI, however, that makes use of gaze as a visual modality which augments speech and elicits joint attention with an artificial agent and which may be used to ground and disambiguate references. Breazeal and colleagues (2005), for instance, provided empirical results using their robot *Leonardo* which showed that people generally use non-verbal behavior such as object-directed gaze to detect errors in the robot's knowledge and to correct these errors. However, the role of intentional states for the occurrence of joint attention in HRI needs to be addressed first in order to establish a link between the utility of gaze in HHI and HCI/HRI.

## 1.4. A Theory of Robot Gaze and (Joint) Attention

The findings on gaze in HCI/HRI reported above are largely subjective measures taken off-line and, in many cases, an observed improvement of the interaction may simply be due to agent/robot gaze behavior engaging the user at a very general level. Psycholinguistic findings (as in Sections 1.1 and 1.2) show, however, that gaze is useful beyond general engagement. Since speaker gaze, for instance, is tightly coupled with her utterance, a listener may use this visual cue to infer the speaker's focus of visual attention and, thus, her referential intentions (Hanna and Brennan, 2007). Therefore, closely observing the partner's gaze during interaction offers benefits for (listener's) utterance comprehension and may also facilitate (speaker's) utterance production (Clark and Krych, 2004). In this thesis, we aim to explore whether utterance-mediated robot gaze can be similarly beneficial for HRI by applying the psycholinguistic findings on speech and gaze production to our robot and observing people's responses to the robot utterances.

As noted above, the use of such language-mediated gaze for joint and shared attention may well be unique to human-human interaction (HHI), possibly relying on (a) a shared biological apparatus and its functions (e.g., eyes that see), (b) certain shared cognitive mechanisms (a person knows from experience that she looks at objects herself, e.g., when mentioning them or when mentioned by others), (c) a theory of mind about our interaction partner (i.e., the ability to reason about *why* someone looks at something) and/or (d) the fact that human gaze is typically informative in some way or another (people almost always look at something or somebody, for some reason). In order to improve robot behavior for HRI, on the one hand, and, on the other hand, to find out to what extent robots are a useful and suitable tool to study human perception and integration of multimodal referential cues, it is essential to examine whether people behave similarly in an HRI setting as in HHI and whether they apply similar expectations and mechanisms in the first place. We provide supporting evidence from five eye-tracking experiments in an HRI scenario, suggesting that people do exactly that.

In these experiments, participants were shown videos of a robot (Figure 1.1) describing objects in a scene while looking at objects. Participants were eye-tracked while observing these videos. They were additionally asked to quickly determine the correctness of the robot's statement with respect to the scene by pressing a button (Experiments 1, 2, 4 and 5), or to correct the robot's false statements orally (Experiment 3). Thus, we consider listeners' eye-movements in the scene, in response to robot gaze and speech, and task responses. Crucially, the tasks that participants were asked to perform in these experiments neither required people to pay attention to robot gaze nor did robot gaze significantly facilitate task completion.

We identify four levels of possible responses when people need to comprehend the robot's spoken statements – accompanied by robot gaze – about the shared environment. Response levels reflect the extent to which people ascribe human-like attentional

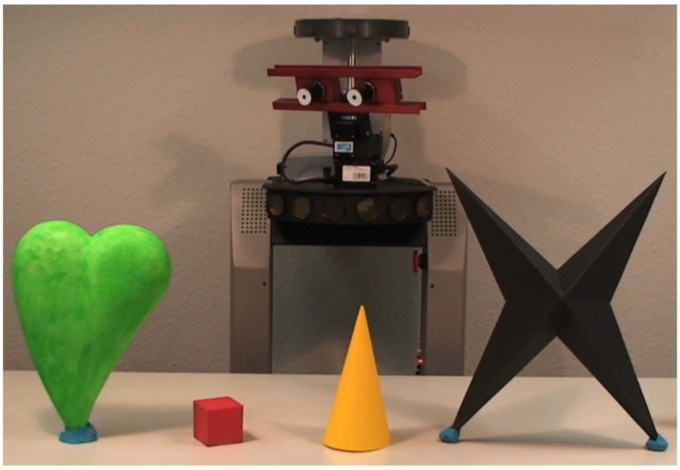


Figure 1.1.: Robot interaction partner. Its head and gaze direction is realized by the stereo-camera mounted on a pan-tilt-unit.

mechanisms and intentions to the robot.

1. People ignore robot gaze if they do not consider the robot to share biological and cognitive mechanism and do not recognize a robot's head/eyes movement as a gaze movement or as a useful cue in general.

2. People may follow robot gaze, possibly reflexively as observed in response to stylized gaze cues (and other symbolic cues such as arrows) in previous studies (Friesen and Kingstone, 1998; Driver et al., 1999; Langton and Bruce, 1999). It is an interesting question whether the visual information obtained after such a reflexive attention shift would affect further (visual and linguistic) processing.

3. If people treat the robot's camera movement as a type of eye gaze – that is, they accept it as a similar way of *seeing* which fulfills similar functions – we predict that people use robot gaze as an attentional cue. Thus, people not only follow robot gaze, they rather seek to find out what the robot attends to and may establish *joint attention* with it. The obtained visual information may be linked (via the robot's attentional state) to the robot's utterance, helping to ground and predict utterance content. Utterance comprehension will, thus, be affected by robot gaze in terms of comprehension speed and/or reference resolution.

4. If people consider the robot to have intentional states, they may further try to

reason about why the robot looks at an object and draw inferences about intentions, i.e. they will try to establish *shared attention* with the robot. Considering that shared attention requires both individuals to consciously and intentionally coordinate mutual attention, it remains an open question whether this can ever be fully established between a human and a robot, in particular when using video-based presentation. However, reasoning about the robot's perceptual and belief states would clearly affect utterance interpretation and inferences people make about the robot's intentions.

The implications for points 3 and 4 are certainly similar and hard to distinguish by purely behavioral observations. Essentially, in level 3 people are assumed to ascribe attentional states to the robot which links gazed at and mentioned objects to infer communicative intentions, while level 4 implies the intentional sharing of perceptual states and goals. That is, people may reason about why the robot did something and what it *intended* to say or do.

Previous studies that were concerned with the assignment of human traits and characteristics to computers (anthropomorphism) and the *mindless application of social rules* in human-computer interaction (Nass and Moon, 2000) mostly considered certain types of language use (e.g., dominant, assertive language versus submissive, equivocal language), outer appearance of the agent or general information about the robot eliciting stereotypical knowledge. While the results of these studies generally encourage the hypothesis that people indeed ascribe intentional states to a robot, our manipulations apply to a visual cue consisting of a simple movement only. Such a cue potentially reveals attentional states, rather than implementing and eliciting social conventions or personality traits, as mentioned above and, yet, may similarly affect utterance comprehension.

## 1.5. Overview of the Thesis

In this thesis we investigate the on-line influence of language-mediated robot gaze on human visual attention, utterance comprehension and intention recognition. We report evidence from five eye-tracking experiments on people's interaction with a robot, exploring whether people use robot gaze to establish joint attention and to draw inferences about the referent intended by the robot. Results from these experiments provide insights on the issue whether people apply similar mechanisms and behaviors when interacting with a robot as they do in HHI. Thus, results from our studies reveal to what extent robots are a useful and suitable tool to further study human perception of gaze. Once it is established that this is a valid paradigm, experiments in such a setting may provide insights on the integration of (possibly conflicting) multimodal referential cues and their interaction. Thus, on one hand, the results from human-robot interaction studies reported in this thesis potentially contribute to the extension of existing theories on the general perception of gaze cues and their effects for interlocutors' attention coordination as well as language processing. On the other hand, implications of the reported work may affect the future development and design of robots, enabling more natural and effective face-to-face communication.

In an extensive review of relevant literature from psychology, psycholinguistics, and human-machine interaction provided in Chapter 2, we present current theories on gaze-processing in human-human interaction and motivate our attempt to replicate and extend some of these findings within a human-robot interaction setting. Furthermore, we consider previous work on the utility of gaze in general human-computer interaction and point out some short-comings that we have tried to overcome in our studies. We additionally motivate our initial experimental design and present results of a pilot study which influenced the design of subsequent experiments. Results from this pilot study were previously published in a workshop paper presented at HRI'08 (Staudte and Crocker, 2008).

In Chapter 3, we introduce the revised experimental design and report findings from Experiments 1 and 2. Specifically, Experiment 1 examined whether human gaze is generally 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) and Experiment 2 determined whether robot gaze is indeed beneficial to the comprehension of robot speech. We manipulated congruency and validity of the produced robot gaze behavior to separate the effects of robot gaze and speech. Results of these experiments relate to response levels 1 (ignoring gaze) and 2 (reflexive gaze-following) identified above. These results have been presented at the Conference on Human-Robot Interaction and were published in the corresponding proceedings (Staudte and Crocker, 2009c).

Experiment 3 is presented in Chapter 4 and served to distinguish between two possible explanations for the results from Experiment 2. Specifically, it was investigated whether people drew inferences about intended referents as a function of robot gaze. Results from this experiment relate to response levels 2 and 3 (joint attention), supporting the hypothesis that people assign intentional states to the robot. These findings have been published in the proceedings of the Annual Meeting of the Cognitive Science

Society (Staudte and Crocker, 2009a,b).

In Chapter 5, Experiments 4 and 5 provide insights on the flexibility of gaze and speech synchronization. Experiment 4 focused on exploring temporal and sequential alignment of robot gaze and speech. While results suggested that temporal synchronization is not essential for robot gaze to facilitate utterance comprehension, a sequentially incoherent order of gaze and speech cues seemed to disrupt people. In Experiment 5, we contrasted this type of coherent and incoherent behavior with neutral robot gaze to study the issue of synchronization in more detail. Results from this chapter provided further evidence in favor of response level 3.

A general discussion of the findings from all experiments and their contribution to current research are provided in Chapter 6. We further discuss ideas for future work before concluding this thesis.

# 2. The Utility of Gaze in Situated Communication

In this chapter, we consider the following kinds and uses of gaze in greater detail.

Section 2.1 reviews social functions of eye gaze which are manifold and comprise, for instance, the expression of emotions, desires and other mental states (Baron-Cohen et al., 1995, 1997b; Adams and Kleck, 2003). The specific production patterns of such gaze cues as well as their interpretation is influenced by social conventions and may therefore vary across cultures, sex or social status (e.g., LaFrance and Mayo, 1976; Dovidio and Ellyson, 1982; Greenbaum, 1985; Yuki et al., 2007; Schofield et al., 2008). The frequency and duration of mutual gaze as a non-verbal cue to help coordinating speaking turns in a conversation may similarly vary with the conventional and personal use of mutual gaze in general (Dovidio and Ellyson, 1982). While a certain pattern for the use of direct gaze cues has been established as relatively reliable (Kendon, 1967), a turn-taking signal can be realized by several cues (Duncan, 1972; Sacks et al., 1974) such that gaze occurrence can vary without immediate function loss (Section 2.2).

In addition to mutual and averted gaze, object-directed gaze also plays a role in social interaction and conveys different kinds of information (see Section 2.3). Object-directed gaze occurs in a pattern that suggests close coupling with the production and comprehension of linguistic content (Allopenna et al., 1998; Meyer et al., 1998; Altmann and Kamide, 1999; Griffin and Bock, 2000). It seems that this type of object-directed or *referential* gaze is indeed produced (and maybe therefore interpreted) in a more automatic manner, probably largely independent of social conventions.

A large body of research reviewed in Section 2.4 supports the view that the way people generally follow gaze (and use it to establish joint attention) is potentially a universal behavior which develops at a very early age (D'Entremont et al., 1997; Moll et al., 2006; Meltzoff and Brooks, 2007) and becomes a reliable source of information in face-to-face communication (e.g., Clark and Krych, 2004; Hanna and Brennan, 2007; Tomasello and Carpenter, 2007). This section further presents evidence indicating what it means to establish joint and shared attention and what processes are involved in these phenomena. Furthermore, gaze-following seems to be such a reliable behavior that previous studies could show that gaze cues elicit *reflexive* attention shifts (Friesen and Kingstone, 1998; Driver et al., 1999; Langton and Bruce, 1999).

In Section 2.5 we explain how referential speaker gaze is used to establish joint attention between speaker and listener and, thus, to facilitate language comprehension for the listener. People follow each other's gaze to jointly attend to an aspect of the environment, possibly because people know that, on one hand, gaze reflects visual inquiry (i.e., *seeing*, Baron-Cohen et al., 1995, 1997b). On the other hand, people have experienced that object-directed gaze is typically closely related to linguistic content so that they use the visual information obtained through gaze-following to infer communicative intentions of their partners. This way, referential gaze and the attentional state it reflects influence further utterance production and comprehension (Clark and Krych, 2004; Hanna and Brennan, 2007).

We present results and insights from previous research on each of the functions of gaze mentioned above and explain how partial results have been applied within the general field of human-computer interaction (HCI) in order to improve communication between robots or virtual agents and human users (Section 2.6). Most implementations of gaze behavior focused on the use of mutual gaze to increasingly convey general engagement of an agent (Kanda et al., 2001; Sidner et al., 2005) or to support the coordination of speaking turns (Cassell et al., 1999a,c; Cassell and Thórisson, 1999). The type of referential gaze that has been observed in spoken human-human interaction and which leads to joint visual attention between interlocutors (see Sections 2.3, 2.4 and 2.5) has to our knowledge not been implemented or empirically investigated within HCI. In Section 2.7 we explain how we have addressed this issue and argue for the validity of our approach.

## 2.1. Social Gaze

The eyes give away reliable information about the emotional and mental state of a person and, thus, play an important role for taking someone's measure. Adams and Kleck (2003) have shown, for instance, that the direction of eye gaze influences what emotion and how quickly this emotion is identified from a depicted face. Participants in this study were given pictures of faces expressing approach or avoidance related emotions (anger versus fear and joy versus sadness, respectively) while showing direct gaze, i.e.

towards the opponent, or averted gaze, that is, away from the opponent. Participants had to indicate which emotions they saw by clicking a mouse button as fast as possible. A second set of stimuli further contained faces with a blend of an emotion pair (anger/fear), again combined with both direct and averted gaze. Response times suggest that gaze direction facilitates the recognition of approach related emotions while averted gaze facilitates avoidance related emotions. Gaze direction further modulated which emotion was recognized in the blended pictures.

While this study suggests that gaze direction and facial expression are not independent of each other, others have shown that eye gaze alone can be a reliable indicator to a person's mental state. Baron-Cohen and colleagues showed that people not only recognize basic emotions but also complex mental states from seeing only the eyes of a person (Baron-Cohen et al., 1997b). In one study, an actress displayed various mental states and emotions and participants were given pictures of either the whole face, only the eyes or only the mouth of the posing actress. A forced choice test where participants had to choose between two words to best describe the picture resulted in very high accuracy for face and eyes display for complex mental states. Accuracy was also very high for recognizing emotions from faces and reasonably high when only the eyes were displayed.

However, the reported results on the interpretation of eye gaze with respect to displayed emotions, attitudes or desires, do not necessarily reveal universally valid gaze patterns. It may be biologically plausible that direct and averted gaze is essentially related to approach and avoidance behavior (Emery, 2000). However, when and how often people directly look at each other, i.e., establish *mutual gaze*, and how emotions are read from eye-gaze can vary considerably. Cross-cultural studies have investigated the frequency and the occasion at which people establish mutual gaze. Yuki et al. (2007), for instance, conducted a study showing that Americans used people's eyes for interpreting displayed emotions to a lesser extent than did Japanese participants. While the former concentrated on the mouth, the latter used mostly the eyes to recognize emotions from both illustrated faces and photographs. One possible explanation according to the authors is that most East Asian cultures like the Japanese require that emotions be frequently subdued. Western cultures, in contrast, appreciate and often demand the overt expression of feelings. This cultural distinction is suggested to promote the eyes for Japanese as an indicator for emotions since eye-movement and the eyes' expression are assumed to be harder to control than muscles around the mouth. Similarly, the mouth is considered more important for people who typically display emotions overtly and rely on the interpretation of those (Yuki et al., 2007). A different aspect of eye-gaze which similarly seems to vary across ethnic and cultural groups concern the use of mutual gaze for interaction regulation (LaFrance and Mayo, 1976; Greenbaum, 1985; Schofield et al., 2008). LaFrance and Mayo (1976) found, for instance, that the frequency of listeners' looks at the speaker varied across race as well as sex. More evidence for differences in (mutual) gaze behavior in gender has been provided by numerous studies, sometimes unexpectedly as shown in an HRI study by Mutlu et al. (2006), or more systematically, reported by Argyle and Dean (1965). The latter further found supporting evidence that the amount of mutual gaze is related to the relationship interlocutors have with each other. Specifically, they showed that the amount of eye-contact is modulated by the physical proximity of interlocutors which is considered to reflect the level of intimacy: The closer two people were placed, the less eye-contact was established – which seems to establish and reflect a certain level of intimacy by itself. It has been shown that the amount of time spent looking at the interlocutor during speaking and listening modulates the impression a person makes with respect to conveyed dominance and social power: Dovidio and Ellyson (1982) varied the proportion of time a confederate spent looking at her interlocutor while speaking conpared to looking while listening. These dyads were recorded and viewed by subjects who rated the perceived social power of the confederate. Results suggested that the more time a person spent looking while speaking (and the less time she spent looking while listening), the more social power was attributed to her. This further shows that social conventions and rules underly certain gaze patterns and how these are encoded and decoded via eye gaze.

## 2.2. Meta-Linguistic Organization of Conversation

Mutual gaze is a cue that people use in face-to-face conversation also to signal and acknowledge whose speaking turn it is (Kendon, 1967; Duncan, 1972). Kendon (1967), for instance, found that a speaker averts her gaze when she begins a long utterance (> 5s duration) and gazes at the listener when she approaches the end of a long utterance. This pattern was established in a study in which Kendon (1967) recorded dyadic face-to-face conversations between participants who had the simple task to get to know each other. Kendon hypothesized that the speaker averts gaze at the beginning of long utterance since planning and execution of the utterance requires concentration and by looking away from the listener the speaker shuts out an additional information source. In contrast, when approaching the end of an utterance, the speaker signals this to the listener and thereby offers her the next speaking turn. This interpretation was further supported by the finding that listeners indeed seem to resume speaking instantly when being looked at and more often fail to do so when not being looked at. While duration and amount of mutual gaze varied considerably between subjects, the use of averted and direct gaze before and after turn boundaries seemed stable. Duncan (1972) further reported and analyzed the use of many more cues for turn-taking such as intonation, body motion and gestures, or pitch and similarly included head turns (considered identical to Kendon's gaze cue). Duncan (1972) showed that when the speaker displayed one or more such turn-yielding cues towards the end of her utterance, the chance of simultaneous turns (when speaker and listener attempt to both talk) decreased dramatically. These results show that meta-linguistic organization is achieved by following rules for verbal and non-verbal signal exchange including mutual and averted gaze. Similarly, efficiency strategies for cognitive processing may influence the usage of such gaze cues.

## 2.3. Gaze in Relation to Linguistic Content

The studies mentioned above have investigated mostly how direct or mutual versus averted gaze reflects and evokes certain mental states. Most of these gaze behaviors play a role in establishing and maintaining a social relationship between interlocutors. Part of this social relationship is the way people coordinate speaking turns (which may, for instance, signal social power or the lack of it) and which can be supported by non-verbal cues such as direct and averted gaze. Another aspect of gaze is related to understanding that eyes capture information about the environment, i.e., that they are used for *seeing* – seeing not only the partner but also objects, other persons, events etc. While mutual gaze potentially provides insights in the interlocutors' social relation or each others intentions and emotions, gaze towards entities in the environment provides the gazer with more visual information about certain aspects of a shared scene. Similarly, a person's gaze towards an object or person also provides the interlocutor with information about what the gazer currently attends to, at least visually. Psycholinguistic research has previously exploited the fact that a person's eye movement typically reflects a shift in visual attention. Observation of such attention shifts provides on-line information about ongoing cognitive processes, for instance, during language production or comprehension (e.g., Tanenhaus et al., 1995; Altmann and Kamide, 2004). Consequently, in the following sections we review relevant findings on the production of gaze during speaking and listening and on its effect on linguistic processing.

### 2.3.1. (Speaker) Gaze during Language Production

Eye gaze promotes a subtle but powerful non-verbal cue that continuously provides on-line visual information about the speaker's visual attention (essentially, gaze is always present, whether produced intentionally or not). The information about where an individual looks complements spoken language and often simultaneously reveals information about an individual's intentions and goals as well as her belief states. Intuitively, gaze cues seem unique among non-verbal cues both in its consistency with internal states of the gazer – a smile may be false, but the eyes often give it away – and its temporal synchronization with the gazer's spoken utterance. This close synchronization of speaker gaze with language production has been established in several studies. It has been shown, for instance, that referential gaze in speech production is associated with the planning process for an intended utterance and typically precedes the onset of the corresponding linguistic reference by approximately 800msec - 1sec. (Meyer et al., 1998; Griffin and Bock, 2000; Griffin, 2001).

Meyer et al. (1998) have conducted two experiments which show that gaze durations are affected by word frequency during a naming task and that this effect is absent when objects have to be categorized. Their participants viewed pairs of line-drawings which they were asked to name. The displayed objects were manipulated with respect to the their contour (complete, contour-deleted) and the frequency of the object names (high, low). Objects with full contours were named faster than those with deleted contours, and objects with high frequency names were named faster than those with low frequency names. Similarly, mean viewing time was shorter for full contour objects and for high frequency objects than for deleted contours and low frequency objects. In a second experiment, Meyer et al. (1998) ruled out that the difference in naming latencies and viewing time was elicited by difficulties during object recognition. Instead of naming objects, participants were asked to categorize objects into existing or non-existing objects by pressing a button accordingly. Object name frequency did not affect decision latencies or viewing times in this second experiment. This indicates that the longer viewing times for low frequency objects arose during lexical retrieval, along with the longer naming latencies, and not during object recognition. Meyer et al. (1998), thus, suggest that people look at an object not only until they have identified it, but further until they have retrieved its phonological form.

Griffin and Bock (2000) further examined the conceptual and linguistic processing involved in apprehending and describing a displayed event. They conducted a study in which participants viewed actions scenes containing line drawings of an event involving two characters, an agent, who is performing an action, and a patient, who is undergoing an action. While one group of participants was asked to describe the scenes during viewing time, a second group first viewed a scene – and prepared their verbal description – before describing it in absence of the picture. A third group had to detect which character was the patient which required comprehension of the causal structure of depicted events, while a fourth group did not have to complete any task at all. Griffin and Bock (2000) manipulated the depicted events such that in one condition scenes elicited predominantly active descriptions. In the second condition, a human character was used as grammatical subject and elicited passive sentences when the human was the patient, and active sentences when she was the agent. In both the speaking-while-viewing and the patient-detection groups, fixations towards patient and agent diverged at approximately 300ms after picture onset. Overt response times, marking patient detection or the beginning of a describing sentence, were also similar in both groups. Both results suggest rapid and complete comprehension of events for both conditions (event comprehension and sentence preparation). Moreover, eye movements of the speaking-while-viewing group revealed that participants inspected a character that they were going to mention approximately 915ms prior to noun onset, regardless of whether the character was subject or object. The manipulation with regard to agent/patient in subject and object position revealed that people generally spent more time fixating the agent before subject onset and less time afterwards. The reverse pattern was observed for the patient, i.e., the patient was looked at longer after the sentences had been started.

Griffin (2001) extended these results by investigating exactly when difficulties in lexical retrieval arise during a spoken sentence, providing insights about the precise time course of word selection during sentence production. An experiment was conducted in which Griffin (2001) asked participants to describe a scene containing three objects using a sentence like *"The A and the B are above the C"*. Griffin found that speakers gazed longer at an object that they were going to name if its name was of low frequency or low codability. A name was considered less codable when it had several similarly dominant names instead of one obvious name. Thus, the duration of referential gaze prior to naming the referent seems to accommodate difficulties of both word selection and phonological encoding. Interestingly, it was also found that the onset of the sentences (as well as viewing time on *A*) varied only according to the frequency and codability of

*A*, i.e., regardless of difficulties related to *B* or *C*. That is, participants started speaking once they had prepared to mention *A* and only later dealt with naming difficulties of *B* and *C*. This result suggests that speakers select their words incrementally.

Taken these results together, they indicate how visual scene information and sentence planning and execution interact. Specifically, the mentioned results indicate that speakers look towards an object before mentioning it as part of a planning process involved in speaking about this object.

### 2.3.2. (Listener) Gaze during Language Comprehension

It has further been shown that listeners' visual attention is driven by the utterances they hear (Cooper, 1974; Tanenhaus et al., 1995; Altmann and Kamide, 1999; Chambers et al., 2004; Knoeferle et al., 2005; Knoeferle and Crocker, 2006). Tanenhaus et al. (1995) found that people *"made informative sequences of eye movements that were closely time-locked to words in the instruction that were relevant to establishing reference"* (p.1632). In one study it was shown, for instance, that people began identifying (visual) referents of a spoken noun already before the offset of the noun. People heard a sentence such as "Pick up the candy" while viewing a visual scene containing a piece of *candy* and sometimes a competing object called a cohort which has a name that shares its onset with the target (e.g., *candle*). Analysis of first fixations showed that people initiated eye movements to the candy before noun offset when the candle was not present. If the candle was present, initiation of eye movements to the target was delayed until or beyond noun offset. Moreover, the findings presented in Tanenhaus et al. (1995) not only show that listeners rapidly fixate mentioned objects, but that the visual context also influences resolution of temporary structural ambiguity in the utterance. Tanenhaus et al. (1995) further reported a study in which participants heard either a (temporarily) structurally ambiguous ("Put the apple on the towel in the box") or an unambiguous sentence ("Put the apple that's on the towel in the box") and saw one of two visual scenes. The one-referent scene contained one possible referent (apple on a towel) and two possible destinations (empty towel, box). The two-referent scene contained an apple on a towel and an apple on a napkin as well as both possible destinations (a towel and a box). Eye movements revealed that in the one-referent visual context people initially interpreted the ambiguous phrase "on the towel" incorrectly as destination whereas in the two-referent visual context the towel was correctly identified as modifier. That is, the phrase "on the towel" was correctly interpreted as modifier since there were two apples in the scene such that the reference would have been ambiguous without the modifier. This result shows that people rapidly integrate visual and linguistic information to comprehend and disambiguate an utterance.

Allopenna et al. (1998) have further established the precise time course with which people look at mentioned referents. In an eye-tracking study, four line-drawings of objects as well as a selection of other shapes were presented on a computer screen to participants. While fixating a central cross, participants were instructed to select one of the objects and move it to a specified location. Besides the target (*beaker*), there was a cohort (*beetle*), a rhyme (*speaker*) and an unrelated object (*carriage*) on the display. Fixations were analyzed from the onset of the target word in the experimenter's instruction. Results showed that people began fixating target and cohort as soon as 200ms after target onset and continued until 400ms after onset. At 300ms after onset, even the rhyme showed an increased probability (though lower then target and cohort) of being fixated. After 400ms, the target then gradually became more likely fixated than the cohort. Considering that programming and launching a saccade takes in itself at least 150-200ms (Matin et al., 1993), these results are indeed evidence for temporally very closely aligned language comprehension and gaze.

Moreover, it has been shown that people not only look at mentioned referents but that they use other disambiguating information from the speech stream which is available prior to the referring noun such as prenominal adjectives (Eberhard et al., 1995; Sedivy et al., 1999) or even verb selectional restrictions (Altmann and Kamide, 1999). Eberhard et al. (1995) reported studies in which participants heard sentences such as "Touch the plain red square". While listening to these sentences, participants saw a visual scene which contained various shapes. In the first condition, there were no other plain shapes such that the adjective "plain" already disambiguated the referring expression. In the second condition, the scene contained a couple of plain shapes of different colors such that the referent could be identified only after the second adjective *red*. In the third condition, the scene contained competing objects that were plain and red but not squares. Thus, the linguistic point of disambiguation was manipulated by the visual context and varied between first and second adjective and the referring noun. Eye movements showed that participants looked at the target before noun onset when the prenominal adjectives already disambiguated the target. Specifically, they looked at a target within 250ms after offset of the disambiguating word. In another study, Sedivy et al. (1999) showed that people similarly process scalar adjectives (as in "the tall glass") and incrementally establish possible referent groups by either contrasting between objects in the visual context (other tall versus short objects) or between an object and a typical representation of that object (a tall glass compared to a typically sized glass).

Altmann and Kamide (1999) further found strong evidence for the rapid use of verb-selectional restrictions during sentence comprehension in the presence of visual scenes. Information provided by the verb elicited anticipatory eye movements to potential referents ('anticipatory' looks to an object occur before it is mentioned explicitly). In two experiments, people listened to spoken utterances while inspecting a visual scene. Participants were first asked to judge whether sentences were valid descriptions of the depicted scenes and, in the second experiment, were not given any particular task. Visual scenes contained several object drawings, e.g., a boy, a cake, a toy train, a car, and a ball. Sentences where the verb indicated only one object as an appropriate referent were contrasted with sentences where verb selectional restrictions allowed all available objects as referents. For restrictive verbs as in the sentence "The boy will eat", anticipatory eye movements to the only edible object in the scene (the cake) were found before noun onset. No such anticipatory eye movements to the cake were found when the verb was unrestrictive and selected several objects (cake, ball, toy train, car) as was the case for sentences like "The boy will move". These findings reveal that, on one hand, linguistic content may be used to rapidly restrict the domain of reference. On the other hand, people's fixations indicate what they consider as potential (visual) referents of the utterance.

Results of a number of studies further showed that visual contexts influence thematic role assignment during sentence comprehension (Knoeferle et al., 2005; Knoeferle and Crocker, 2006). In three experiments, Knoeferle et al. (2005) investigated the comprehension of (preferred) subject-verb-object (SVO) sentences and (less preferred) object-verb-subject (OVS) sentences in the context of depicted events. The sentences described depicted events which contained a role-ambiguous character (e.g. a princess), acting and being acted upon, as well as an agent character (e.g., a fencer) and a patient character (e.g., a pirate), such that the fencer paints the princess and the princess washes the pirate. Sentences were temporarily role-ambiguous since the first noun phrase referred to the role-ambiguous character (princess) – and there were no case-marking cues to determine the correct syntactic and thematic relations (nominal and accusative feminine articles are identical in German). The second noun phrase, which was unambiguously case-marked as subject or object, disambiguated the sentence structure and role assignment. For early disambiguation, listeners had to rely on depicted event scenes that showed fencer, princess and pirate. Thus, as soon as the verb identified, for instance, the washing action, eye movements indicated that participants had established the princess as agent of the event (SVO) and not patient (OVS). That is, fixations on the character involved in the depicted event (pirate as patient of the washing action) increased which reveals rapid integration of depicted and linguistic information to assign sentence structure, thematic roles and to anticipate the next referent.

Summarizing the mentioned results, it seems that gaze is not only driven by what is heard, it also serves to continuously gather visual information which is integrated into the comprehension process and which may clearly affect interpretation of the unfolding utterance.

## 2.4. Joint and Shared Attention

In the previous section we have presented findings on gaze production during speaking and listening and what this gaze reveals about language processing. These findings are limited to utterance-mediated gaze, i.e., gaze that is mainly driven by what is said or heard. In face-to-face communication, interlocutors not only speak to each other, they can further see and use each other's gaze. Essentially, people use another person's gaze *because* they understand that eye gaze towards an entity in the external world reflects visual perception of that entity, and because they understand that there may be reason for this other person's gaze, e.g., interest, danger or food (Baron-Cohen, 1995; Emery, 2000; Flom et al., 2007). Thus, it can be useful for person A to know what person B (visually) attends to since the object in question may be interesting or dangerous or in some other way relevant to person A as well. Following B's gaze to an object reveals potentially interesting information for A and results in *joint attention*, a state where both individuals end up attending to the same object, or even *shared attention*, a state where partners are aware of each others attentive state and draw inferences about each others intentions from this.

We follow Emery (2000) in considering *joint attention* to occur when a subject follows another subject's gaze to mutually attend to an entity. Joint attention presupposes that the gazer has attentional states such that the follower has reason to consider the looked-at entity as relevant. The term *shared attention* is used to refer to a similar phenomenon which additionally involves intention sharing: One person *intentionally* directs another person's gaze to an object by looking at this object, in order to communicate goals for cooperating in task completion or just to share the experience (Emery, 2000). Notably, what we call shared attention has been named joint attention previously (Kaplan and

Hafner, 2006; Tomasello and Carpenter, 2007). Similarly, it has been described as a state that requires that the "goal of each agent is to attend to the same aspect of the environment" (Kaplan and Hafner, 2006, p.144) and that "both agents are aware of this coordination of 'perspectives' towards the world" (Kaplan and Hafner, 2006, p.145). However, we decided to adopt a more fine-grained categorization and distinguish joint and shared attention.

The ability to follow gaze, indicating the perception of others as beings with attentional states, develops already in infants. Previous research on infant perception of eyes and head direction has shown that children learn to perceive gaze as meaningful and potentially revealing something new, at a very early stage in development. The age at which infants first follow gaze is controversial due to different methodologies (e.g., experimenter versus infant's mother as gazer/interlocutor, angle of produced gaze movements) or different definitions of gaze (e.g., including head turns or not). However, D'Entremont et al. (1997) have shown that infants, even as young as three months old, already follow a person's head turns towards a puppet.

Support for a conceptual distinction of joint and shared attention may be further drawn from the developmental stages of infants. Meltzoff and Brooks (2007) suggest, for instance, that infants at the age of 10 to 12 months are at a transitional age, capable of gaze-following but not of intentionally sharing experience. They showed that at this particular age infants follow a person with open eyes and refrain from doing so when the person's eyes are closed. While they realize that closed eyes do not signal visual attention, they do not understand that blindfolds similarly obstruct vision. And yet, the infant must have understood that the gazer has a seeing organ – just like herself – that indicates what the gazer (visually) attends to and towards which the infant then follows. It seems that children at this stage establish joint attention but not shared attention, i.e., they do not fully grasp that they can direct and share the gazer's view intentionally. Thus, the authors suggest that only between 12 and 18 months of age infants learn to share their interlocutor's view and interpret it as an indicator to her goals and intentions. This is in line with findings from Moll et al. (2006) who showed that 14-month-olds are able to reason about what an adult most likely attends to given the adult's gaze direction and her past experience.

The presented studies suggest that children typically learn to establish first joint and later shared attention at an early developmental stage. Many autistic children also begin to follow gaze and head turns towards objects (Leekam et al., 1998; Kylliäinen and Hietanen, 2004). Even though with a certain delay (Leekam et al., 1998, 2000), they also learn to follow gaze. However, autistic children seem to learn to establish only limited joint attention and no shared attention. That is, they do not infer intentions from the perceived gaze direction nor do they infer goals, desire or even interest in the object in focus. Results from a number of studies involving perception of gaze in normal and autistic children suggest that the observed inability to read mental states such as desire, intention or interest from a person's gaze is indicative of an inability to form a theory of mind (Baron-Cohen et al., 1995, 1997a,b). This is further evidence suggesting that people normally follow gaze and interpret it with respect to mental states because they ascribe attentional and intentional states to the gazer and because they seek to establish joint and shared attention with the gazer. Consequently, gaze is an essential cue in the context of studying phenomena such as joint and shared attention, both in HHI and HRI.

Summarizing this section, the insights from developmental studies on the role of gaze, also for autistic children, further support the distinction of joint and shared attention. This suggests that basic joint attention requires the general ability to follow and understand object-directed gaze (i.e., to interpret it as an attentional state which may reveal communicative intentions) which children learn very early. The ability to reason about the goals behind a gaze cue seems to require that the interlocutor assigns intentional states to her communication partner, i.e., that she has a theory of mind for her partner. Shared attention further requires the understanding that the partner's attention can be manipulated by one's own gaze, thus, also manipulating what the partner believes about oneself.

### 2.4.1. Reflexive Gaze-Following

Related research has further suggested that gaze-following is a behavior that is indeed applied so reliably that it may be considered automatic. Specifically, previous studies have shown that people *reflexively* follow gaze cues and also other direction-giving cues such as arrows (Langton et al., 2000; Ristic et al., 2002). It is an ongoing debate whether attending and reacting to eyes and gaze is "hard-wired" (Baron-Cohen et al., 1997b, p.328) in the sense that it is a unique attentional process with a dedicated neural basis (Baron-Cohen et al., 1997a; Emery, 2000) or whether the immediate, low-level and reflexive attention shift that gaze cues elicit (Friesen and Kingstone, 1998; Driver et al., 1999; Langton and Bruce, 1999; Vecera and Rizzo, 2006) similarly applies to other attention directing cues such as arrows (Bayliss and Tipper, 2005; Tipples, 2008). While reflexive attention shift (especially in the context of peripheral cueing) is considered to be *exogeneous*, voluntary orienting towards a symbolic cue is called *endogenous* (Posner, 1980). Studies suggesting that gaze cues reflexively trigger (exogenous) attention shifts have typically presented stylized faces or eyes (or arrows) to their participants who then had to detect or identify a target stimulus that appeared either in the cued or uncued direction. Findings revealed that response times where significantly shorter for the cued location when the target stimulus appeared within a certain time window (stimulus-onset asynchrony, SOA) after the cue: 1005ms in Friesen and Kingstone (1998), 1000ms in Langton and Bruce (1999) and 700ms in Driver et al. (1999). Crucially, these cueing effects were observed even though cues were not predicting the target location, i.e., were uninformative. In addition to these early cueing effects ascribed to reflexive orienting, it has been shown that these cues can also trigger voluntary attention shifts when they predicted the target location (Friesen et al., 2004; Tipples, 2008). In their study, Friesen et al. (2004) used counterpredictive cues which predicted the target location in the location opposite to the cued location. It was found that people initially attended to the cued location (cueing effect for short SOAs, up to 600ms) but then also attended to the opposite location in which they predicted the target to occur (cueing effect for longer SOAs, from 600ms to at least 1800ms). This seems to suggest that involuntary and voluntary use of cues are separable processes. However, recent studies which correlate voluntary and involuntary attention shifts have questioned this assumption. Tipples (2008) presented evidence showing that what appears as involuntary or reflexive orienting is at least influenced by voluntary attentional control. People that scored high in an attentional control questionnaire (i.e., who reported "good" attentional control) also showed larger involuntary orienting effects. Moreover, Vecera and Rizzo (2006) presented a study on neural impairment and attention from which they conclude that gaze triggers the type of voluntary attention shift that is also observed for words, for instance. Thus, it seems that reflexive and voluntary attention shifts cannot be entirely decoupled and rather both determine when and where an individual shifts her visual attention.

Notably, all above mentioned studies relied on the presentation of static cues even though gaze is typically a dynamic cue. Additionally, these studies typically did not involve recording people's overt visual attention shifts (eye movements) and restricted themselves to reaction time for measuring detection or identification time. Importantly, experiments within this paradigm looked mostly at visual orienting in response to a simple visual stimuli and did not consider the interaction of language – or, more generally, intentions potentially involved when considering "real" gaze – with these gaze cues. As a study by Hietanen et al. (2008) suggests, pictures of faces do not necessarily have the same effect on an observer as a real face does. The observer may lack the feeling of being looked at since she does not attribute any intentional, social meaning to the gaze cue.

The described studies do suggest that gaze can elicit both levels of response, reflexive as well as voluntary orienting, which raises the question about how these may co-occur and possibly interact. Moreover, it is unclear whether it is also voluntary orienting when children, for instance, follow their mothers' gaze to establish joint and shared attention. Previously, attention shifts have been called voluntary or volitional in the context of (covert) orienting when such an attention shift was elicited by a central symbolic cue that predicted a target in an uncued location (e.g. Friesen et al., 2004). There may be a qualitative difference between using such a symbolic cue (after being told that it is useful and having trained to interpret it accordingly) compared to following someone's gaze to an object or person and inferring mental states of the gazer (because as a child one has learned that gaze-following potentially reveals interesting information).

## 2.5. Joint Attention and Language Comprehension

In the previous sections, we have explained that an important aspect of gaze is related to understanding that eyes capture information about the environment. Knowing that an individual's gaze is often directed to entities in the vicinity and that this provides the individual with (visual) information about this entity makes gaze-following a useful strategy for learning (what does an unknown word refer to), survival (is there a source of danger) and smooth communication (what is my partner going to say, want or do). Baron-Cohen and colleagues (1995; 1997a; 1997b) showed in a number of studies, for instance, that a speakers' gaze direction can normally be a significant cue to the intended referent of the speaker. In one study, children were shown two nonsense shapes and were asked to indicate which of them was *beb*, a nonsense word. While first they had to guess and deliberately pointed at one shape, the second time a cartoon face named Charlie was placed between the shapes and looked at one of the shapes. Asking the children what Charlie thought was the *beb*, most of them pointed to the one that the face was looking at. Children with autism, in contrast, mostly stayed with their initial decision and failed to interpret the face's gaze cue as an indicator for attention and desire with respect to a certain shape. These studies by Baron-Cohen and colleagues seem to suggest that autistic children typically do not read mental states from the eyes at all and even tend to prefer artificial cues such as arrows over reading eye direction. These results also suggest that gaze is an important cue to an individuals intentions and that not being able to interpret it as such indicates a deficiency in theory of mind formation which further disrupts social interaction (*"it is the lack of mental state concepts that causes the failure to understand that eye-direction signifies this range of mental states"*, Baron-Cohen et al. 1995, p.394).

In addition to this general notion of visual attention and intention ascribed to gaze, a close coupling has been established between produced gaze and language comprehension and production (reviewed in Section 2.3). Whether, and precisely how, the close alignment of gaze with spoken language production, for instance, helps listeners to identify and anticipate utterance content, is subject to ongoing research. The mentioned studies on joint and shared attention, however, clearly suggest that people do monitor and use each others gaze in face-to-face communication. In spoken communication, information obtained through gaze-following helps to rapidly ground and resolve spoken utterances with respect to a common environment (Moore and Dunham, 1995; Clark and Krych, 2004; Tomasello and Carpenter, 2007). Speakers' gaze to an object can, thus, function as a visual reference to an object, augmenting linguistic references. Consequently, face-to-face communication produces not only utterance-mediated gaze, but also gaze-mediated gaze which potentially reflects states of joint visual attention.

Studies investigating the utility of such referential gaze cues in face-to-face communication have provided evidence that listeners use speakers' gaze to identify a referent in the scene before the utterance unambiguously identifies that referent (Hanna and Brennan, 2007). In a first experiment, Hanna and Brennan (2007) found that listeners follow and use speaker gaze to constrain their domain of interpretation such that (a) temporary ambiguity is disambiguated, and (b) reference resolution is enhanced since this information is available early during language processing. The experiment was conducted with a director and a matcher facing each other. Both had their own displays hidden behind a low barrier but were shown the other's display at the beginning of the experiment. Displays contained either a mirrored object constellation, i.e., were congruent with each other as shown in Figure 2.1, or contained different spatial object arrangements (non-congruent) such that the director's gaze was uninformative. The director instructed the matcher to move one of the displayed objects to a specific location. Such an instruction contained a referring expression of the form "the [color]

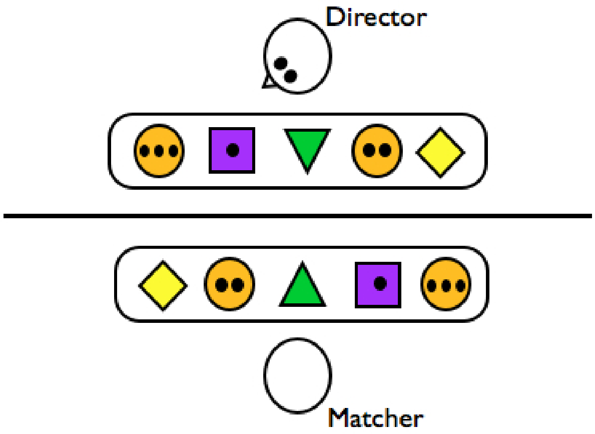


Figure 2.1.: Sketch of experimental setting with reversed displays containing a target ("orange circle with three dots on it") and a far competitor, as described in Hanna and Brennan (2007). Original pictures and a more precise description can be found in the respective paper.

[shape] with [number of dots]". The display either contained a competitor object of the same shape and color next to the target ('near competitor') or further away ('far competitor') such that the referring expression was temporarily ambiguous, or it contained no competitor. In the 'near competitor' condition the director's gaze towards the target was not clearly disambiguating while in condition 'far competitor' director's gaze more clearly distinguished between target and competitor. Results from matchers' target looks showed that the matcher identified the target before the linguistic point of disambiguation if displays were congruent. Moreover, in the 'far competitor' condition participants seemed to identify the target as early as when there was no competitor at all, suggesting that director's gaze was clearly disambiguating.

In a second experiment, Hanna and Brennan (2007) changed the display arrangements such that displays were either congruent (mirrored) or reversed. In the reverse condition, objects on the director's right were to the matcher's right such that director's gaze needed to be re-mapped in order to be informative from the matcher's perspective. Matchers' target fixations indicated that matchers used directors' early target fixations (visual point of disambiguation) to initially orient towards the same (mirrored) side of their display. 1000ms after the visual point of disambiguation, matchers apparently remembered the display condition and began to adjust to that, i.e., oriented towards the opposite, "target" side when displays were reversed. The congruent condition replicated previous results showing that director's (i.e., *speaker's*) gaze is an early disambiguating cue. Though matchers did not immediately follow directors' gaze cue to the target they seemed to use this information about 1500ms later, between the color onset and the linguistic point of disambiguation, to identify the target. Interestingly, matchers' target fixations in the reverse condition showed a considerably smaller but nonetheless significant benefit of director gaze for target identification. This suggests that speaker gaze helped to identify referents even when this gaze cue was initially misleading. Listeners seemed to establish a mapping of the speaker's gaze to their own visual scene and, still, made use of the speaker's gaze early during comprehension.

The results from Hanna and Brennan in addition to other previous results (e.g., Baron-Cohen et al., 1997b), suggest that people infer intended referents from the speaker's gaze after the initial, reflexive response to gaze (Friesen and Kingstone, 1998; Driver et al., 1999; Langton and Bruce, 1999). That is, beyond the possibly reflexive attention shift in response to gaze, people seem to be able to impose the communicative context onto the visual stimulus and, thus, may still interpret the gaze cue as a visual reference which reflects communicative intentions.

The above mentioned findings show that gaze during spoken communication is systematically and automatically coupled to situated speech. This close coupling in addition to the general notion of *seeing* and visual attention ascribed to gaze may be the reason that listeners interpret speakers' eye movements on-line as visual references to help rapidly identify, and disambiguate among, intended referents.

## 2.6. Social Robot Gaze

In previous sections, we have reviewed in detail the role of gaze in human-human interaction (HHI). The reported results have highlighted the utility of gaze for rich and smooth interaction between individuals. It was shown that gaze is closely coupled to speech and how the partner's gaze reveals what she understands and plans to say. Further, it was explained that gaze is a cue that guides visual attention reflexively and voluntarily and that already infants learn to use gaze cues for further information processing. Considering that gaze is such a frequent, reliable and useful cue, it is conceivable that it also plays a role when humans interact with virtual agents or robots. To investigate to what extent the findings on human gaze are also valid for human-computer/human-robot-interaction (HCI/HRI), previous research in this area has considered human gaze patterns and their application to agents. User studies from HCI, thus, involve various functions of gaze as, for instance, non-verbally engaging the user, turn coordination, or simply making the interaction more pleasant (Sidner et al., 2005; Cassell et al., 1999a; Cassell and Thórisson, 1999; Kanda et al., 2001).

While agent gaze has been shown to enrich HCI in terms of "enjoyment" (Kanda et al., 2001), it has further been shown to negatively influence people's judgement of the robot's competence and reliability (Sidner et al., 2005). Sidner et al. (2005) conducted a study investigating how robot gaze affects people's impression of the robot as well as their own non-verbal engagement in a conversation. Participants were asked to interact with the robot *Mel* and then rate, for instance, their liking of the robot, their sense of involvement, and their impression of reliability. Participants in the first group interacted with Mel when it produced both verbal and non-verbal behavior (mover condition) while a second group had to communicate with Mel when it produced speech only (talker condition). In the mover condition, Mel was capable of looking and pointing at objects when the task explicitly required it to draw the user's attention to an object. That is, Mel turned towards the table top between the interlocutor and itself after it had explicitly referred to an object on the table top and needed to make sure the partner had seen it as well. Furthermore, it looked towards its partner whenever it finished a speaking turn. User ratings revealed that neither participants' liking of the robot nor their factual knowledge gained during the interaction were affected by the conditions. In contrast, people's judgement of the robot's reliability was strongly affected by Mel's non-verbal behavior. That is, when Mel performed no movements it was rated to be *more reliable* than when it used head/gaze movements and pointing gestures. One explanation for this result may be that the produced head/gaze movements were simply not appropriate and instead enforced the perception of incompetence of the robot. However, Sidner also found that participants in the mover condition felt more involved in the conversation than participants in the talker condition. These participants also showed more non-verbal behavior themselves in this condition (e.g. more mutual gazes).

A point of criticism that this study has in common with similar HCI experiments (see also Kanda et al., 2001; Wang et al., 2006) is that they lack an appropriate baseline condition. To assess the friendliness and enjoyability of a more or less plausible gaze pattern, the robot shows such a pattern in one condition while in the baseline condition the robot shows no gaze movement at all. It is difficult though to evaluate the actual contribution of the implemented gaze pattern when compared to a system without any command of the gaze modality, especially considering that gaze possibly comes at a cost such as decreasing the impression of competence and reliability (Kanda et al., 2001; Sidner et al., 2005).

Moreover, it has been attempted to employ gaze cues for implementing social behaviors mentioned in section 2.1. Specifically, Kipp and Gebhard (2008) have shown in a study with a virtual character that direct gaze can be used to manipulate the impression of social dominance (of the virtual character as perceived by participants). While continuously direct gaze (which has been dubbed the *Mona Lisa* strategy) conveys dominance, direct gaze during speaking combined with averted gaze during listening also conveys dominance but is perceived as more negative and close to arrogant. These results have essentially replicated the effects of direct gaze during speaking and listening reported for HHI by Dovidio and Ellyson (1982).

Cassell et al. (1999c) took a different approach to implementing *natural* gaze behavior. As the literature suggests, mutual gaze is a signal that is used to coordinate turn-taking in a conversation (Kendon, 1967; Duncan, 1972; Sacks et al., 1974). The authors hypothesized that gaze also correlates with information structure of the discourse, that is, the theme (what is known; links the utterance to previous discourse) and the rheme (new information) of an utterance. Initial experiments confirmed the correlation of speaker gaze towards and away from the hearer with both turn-taking and information structure (Cassell et al., 1999c). In this study, participants who were strangers to one another were told to sustain a conversation on any topic for at least 20 minutes. Three such dyads were video-taped and transcribed in terms of speech, speaker gaze towards and away from the listener, and head nods. Speech was annotated using the units turn, rheme and theme. A beginning turn was defined as the first word of a new turn, and the end of turn was defined as the last and second last word of a turn. Theme and rheme boundaries were similarly defined. The analysis revealed co-occurrences of beginning direct and averted gazes with turn beginnings and endings as well as rheme and theme beginnings. It was confirmed that speakers often (but not always, 44% of all turn beginnings) look away from the listener upon turn beginning. Interestingly, *all* turn beginnings that also began a theme co-occurred with a look away. Similarly, it was found that speakers look towards the listener at the end of a turn, but only in 15% of the turn endings. However, they looked towards the listener at *all* turn endings that co-occurred with a beginning rheme.

Cassell and her colleagues used these results and implemented a heuristic for gaze production. Original implementations included the realization of turn-taking cues, only, and were tested on embodied conversational agents (e.g. *Rea* and *Gandalf*) in interaction with users (Cassell et al., 1999a; Cassell and Thórisson, 1999; Cassell et al., 1999b). In such a user study, participants were asked to interact with three different characters within the *Ymir* environment (Thórisson, 1999). These characters differed with respect to the non-verbal feedback they gave: No non-verbal feedback, emotional feedback containing of smiles and confused expressions, and envelope feedback comprising turn-relevant gazes (Cassell and Thórisson, 1999). Analysis of users' verbal contributions to the interactions revealed that people made fewer contributions in the third condition. This result was considered to show that conversation was more efficient when the character used gaze behavior to indicate turn endings and beginnings. Subject ratings further showed that participants judged a character's language abilities as well as 'interaction smoothness' to be higher when it used turn-relevant gaze behavior than when it did not. It is unclear, however, what ratings for 'smoothness of interaction compared to interacting with a dog' are meant to reveal.

More recently, Mutlu and colleagues implemented the initial probabilistic algorithm suggested by Cassell et al. drawing on both turn-taking and information structure effects on gaze production. This implementation was used and evaluated on a storytelling humanoid robot (Mutlu et al., 2006). The authors manipulated the probabilities in the algorithm such that the robot produced two versions of gaze behavior differing in the overall amount of looks towards each of two listeners. The generated robot behavior was evaluated by measuring participants' performance at recall of the heard story and by giving out pre- and post-experiment questionnaires. The results showed that participants who were looked at by the robot more often (at appropriate occasions) performed better on the recall task. While impressions of the robot with frequent mutual gaze behavior were again not all positive, the effect on recall performance suggests that people attend closer to the robot when being looked at more often.

Similarly, Yamazaki et al. (2008) have shown that robot gaze towards the listener at turn endings could be useful in interaction with a human. Specifically, it has been shown that direct robot gaze at turn end elicits more head turns and nods from the subject than direct robot gaze at the beginning or in the middle of a turn. This result indicates that participant's non-verbal engagement depends not only on the robot's production of non-verbal behavior in general (compared to no production thereof) but also on *when* the robot issues head/gaze turns.

Notably, the reported studies have failed to directly investigate whether robot gaze, similar to human gaze, conveys attentional states such that it elicits joint attention between robot and user – and whether this helps the user to resolve ambiguous references by means of visually constraining the domain of reference. There have been few attempts in HCI to employ robot gaze as a modality that expresses attentional states of the robot and potentially elicits joint attention with the user such that ambiguous references, for instance, can be resolved. The work conducted by Breazeal and colleagues with the robot *Leonardo* (Breazeal et al., 2005) has made relevant contributions to this field of research. It was shown that implicit robot behavior like gaze shifts, head nods and other gestures was used by human interlocutors to faster solve a collaboration task. Specifically, participants were asked to interact with Leonardo and make it switch on buttons that were located in front of it. There were two conditions, an implicit and an explicit one. In the explicit condition, Leonardo looked at buttons right before it was going to press it or point to the button itself. In the implicit condition, it additionally looked at buttons when the interlocutor pointed at them, and it produced general gaze shifts and eye-blinks to convey liveliness and shrugging gestures to convey confusion.

A post-experiment questionnaire revealed that participants in the implicit condition thought they had a better mental model of the robot (they could tell when the robot was confused or had understood what was referred to) than participants in the explicit condition. Furthermore, some behavioral data was analyzed such as number of errors and repairs in a conversation or time needed to complete the task (make Leonardo turn all buttons on). Results revealed that the implicit, non-verbal information helped people to detect errors in the robot's performance and, consequently, to repair them. Not surprisingly, task completion time was considerably shorter in the implicit condition than in the explicit condition, in particular, when errors occurred throughout the conversation since their detection was facilitated. Breazeal et al. (2005) concluded that Leonardo's gaze constituted a "window to its visual awareness" and that people perceive the robot's gaze-following as signaling "shared attention" (Breazeal et al., 2005, p.714).

While the described studies have suggested that non-verbal behavior in general, and gaze in particular, influences the way people perceive and interact with an agent or robot, they do not reveal precisely how referential gaze influences utterance comprehension, for instance, and whether people infer intentional states from the robot's gaze. Nevertheless, the presented findings are promising and suggest that people might indeed interpret gaze cues with respect to attentional and 'mental' states and that appropriate robot gaze can facilitate interaction.

### 2.6.1. Appearance and Motion of Inanimate Entities

When working with robots and other agents, naturally the question arises at some point to what extent (only) the appearance of the agent influences people's perception thereof. Previous studies, partly described in the previous section, have hinted at an important role of the agent/robot form suggesting that people expect a humanoid form to also reflect human-like behavior (Kiesler and Goetz, 2002; Hegel et al., 2008; Groom et al., 2009). Previous studies have shown, however, that also motion patterns can reliably convey goal-directness and intentional states of things which otherwise do not look human-like or in any other way intelligent (Heider and Simmel, 1944). We, thus, argue that the appearance of the robot's head/eyes is not crucial to participants' responses: Firstly, robot head and gaze do not necessarily have to be distinct modalities as head direction is a cue similar to gaze, often used when eyes are obstructed or when visual attention needs to be directed to other issues (Emery, 2000; Imai et al., 2002; Hanna and Brennan, 2007). Secondly, people merely need to understand that the robot's camera is their 'organ' for *seeing* in order to assign meaning to its gaze (see e.g. Meltzoff and Brooks, 2007; Baron-Cohen et al., 1995). It seems that this understanding does not rely on the human-like appearance of the seeing-organ (the eyes) but that this rather results from a plausible movement pattern and the conveyed functionality of the camera. How powerful motion can be in conveying intentional, or at least goal-directed behavior has been impressively illustrated by an animated film (Heider and Simmel, 1944), also used for further experimental research by, for instance, Berry et al. (1992). Heider and Simmel (1944) produced a simple animation showing moving geometrical figures: a small and a large triangle, a disc or circle, and a large rectangle with a section that opened and closed like a door. The animation showed movement of various types: (a) Successive movements with momentary contact of two shapes, (b) simultaneous movements with prolonged contact, (c) simultaneous movement without contact, and (d) successive movements without contact. Participants were asked to describe what they saw and typically interpreted movements as actions of animate beings, in fact mostly as those of persons. Heider and Simmel (1944) found that movements of type (a) were often interpreted as one object *hitting* the other, (b)-movements were considered as *pushing* or *pulling* actions, (c) often as *leading* or *chasing* events, and (d) as *chasing* or *evading*. The interpretation of the events (e.g., chase versus lead or flee) depended on and reflected what people considered to be the origin of the movement. Besides the origin of movement, people seem to further interpret motives for movement. Instead of using the term "entering", for instance, when a small triangle moves into the large square, it has been called "hiding" or "being-forced-in", suggesting that they ascribed intentions and motivations to the moving objects.

## 2.7. Studying Robot Gaze as an Attentional Cue

In this section, we draw on the findings in the various research areas reported above to motivate our own investigation of the role of referential speaker gaze in situated interaction with a robot. Considering the findings described in Sections 2.3 and 2.5, we 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 human gaze production patterns, A says "Pass me the mug, please." and looks at the mug 800-1000ms before saying "mug". As part of understanding A's utterances, listener B then looks at the mug around 200-300ms after A started saying "mug". This would result in a 1000-1300ms 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 mentioning of the mug. The time span between A's and B's gaze towards the mug is shortened dramatically and B can rapidly ground A's reference to the mug in their shared environment. Also B's looks to the mug, rapidly inform A that the utterance has been understood. Furthermore, in a situation where there are several mugs, gaze may provide a crucial means of referential disambiguation.

A human-robot interaction scenario offers a controlled setting for the manipulation of gaze parameters, such as temporal alignment of gaze and speech, as well as the observation of people's responsive behavior. Thus, the scenario described above can be conceived also as an experimental HRI setting with a robot as speaker A, for instance. Despite a general and large interest in using gaze cues also in HCI/HRI, the psycholinguistic findings on gaze as referential cue described above have not yet been examined empirically. Thus, we investigate whether referential robot gaze is a useful cue for people to disambiguate the robot utterance, to infer the robot's referential "intentions" and whether this can be evaluated by means of on-line, quantitative measures. Such measures are crucial when investigating issues in information processing that people may not be aware of and, thus, cannot report accurately in questionnaires, for instance. Moreover, we argue that the insights gained in such a scenario can be compared to HHI behavior such that these results potentially contribute more broadly to the investigation of human gaze processing and joint attention.

That is, we propose an experimental setting in this thesis, involving a robot speaker and a human listener. We apply previous results from psycholinguistics in order to produce cognitively motivated robot behavior. Specifically, the robot's gaze is timed such that the robot fixates an object it is going to mention around one second before the corresponding noun onset (Meyer et al., 1998; Griffin and Bock, 2000; Griffin, 2001). This behavior is then video-taped and played back to participants. If such human-like gaze behavior elicits natural responsive behavior from human interlocutors, we expect to observe 'speech-following' (Tanenhaus et al., 1995; Altmann and Kamide, 1999; Sedivy et al., 1999; Knoeferle et al., 2005) as well as gaze-following behavior (Hanna and Brennan, 2007; Meltzoff and Brooks, 2007) that is typical for HHI.

Although it might be argued that video-based presentation of the robot does not allow 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 (Woods et al., 2006). Further, the subjective perception of remote versus collocated agents (for both robots and virtual agents) has been studied by Kiesler et al. (2008) and similar results were presented. One might further argue that using a virtual agent instead of a robot could solve this problem of video presentation. However, even though there exists a growing body of research on gaze for both virtual agents and robots simultaneously, there is one main difference between the two types of agents that potentially affects the usage of gaze. While an agent "lives" in its own world and is assumed to have complete knowledge of its environment, a robot shares the environment with a human interlocutor and is not expected to have full knowledge of the world. This leads to different expectations and impressions of an agent versus a robot. Mistakes and errors are potentially more acceptable and less irritating when communicating with a robot. Moreover, robots seem to elicit more anthropomorphic interaction and attributions than agents (Kiesler et al., 2008). Kiesler and colleagues also investigated whether a robot and an agent in co-present and remote conditions are perceived differently. The results of their questionnaire, which participants had to fill out after a 10-15min discussion with the robot/agent, indicate that the robot was perceived as more life-like, having more positive personality traits and being liked better. However, whether the robot/agent was co-present or remote (i.e. recorded and projected onto a screen) did not seem to greatly influence participants' impressions. This study, thus, supports our decision to employ a robot for studying the benefits of cognitively-motivated, referential gaze behavior and suggests that remote, video-based presentation should not substantially affect perception or interaction.

Further, if the participant behavior observed within our HRI setting is indeed similar to people's behavior in HHI, we argue that such an experimental design may be useful to shed light not only on the role of robot gaze but on how humans process gaze in general. We intend to approach questions like: Can we measure the utility of referential gaze, i.e., its information content and benefit for utterance comprehension? How are potentially mismatching linguistic and gaze cues integrated? As Hanna and Brennan (2007) already showed, listeners can map and use speaker gaze for disambiguation. That is, listeners infer the intended referent from the speaker's gaze in order to anticipate and quickly resolve the upcoming referring expression – even when listeners have to take into account that their own and the speaker's visual scene differ.

In the experiments presented in subsequent chapters, we examine whether listeners similarly use information obtained from robot gaze in order to infer referential intentions and facilitate comprehension. The general setting of these experiments was 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. For the production of robot gaze behavior, we made use of the psycholinguistic findings summarized in Section 2.7. That is, for producing referential and cognitively plausible robot gaze, the camera moved towards an object approximately one second prior to its mention, which is consistent with the observed co-occurrence of referential gaze and referring expressions in human speech production. Participants were typically instructed to attend to and determine the 'correctness' of robot utterances with respect to the scene. We consider two dependent measures: People's eye movements were monitored as an on-line measure of visual attention, and people's responses to different tasks were used as off-line measures, e.g., of the effort that comprehension requires. We chose such a video-based presentation of the robot in order to better control experimental conditions and to obtain statistically relevant data. 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 et al. (2008); Woods et al. (2006).

Using the outlined setting, we investigated in Experiment 1 whether people attend to robot gaze as they typically attend to human gaze. That is, participants are shown videos of a robot that looks at an object and subsequently mentions a referring expression. Since participants are asked to validate the robot utterance, they need to resolve the referring expressions with respect to the scene. Participant's eye movements are expected to reveal whether people follow robot gaze to an object and, further, whether they use this visual cue to resolve the referring expression with respect to the looked at object, even when there are other possible referents in the scene. Results from this experiment relate to response levels 1 and 2 described in Section 1.4 and indicate whether people (can) ignore robot gaze or whether it generally directs people's visual attention.

Experiment 2 sought to study the influence of robot gaze on utterance comprehension in greater detail. We manipulated robot gaze congruency to explore whether robot gaze is interpreted as a visual reference revealing the robot's attentional states such that congruent gaze would facilitate comprehension while incongruent gaze might mislead people and, thus, disrupt comprehension. Findings from this experiment provided further evidence for the hypothesis that people follow robot gaze (supporting response level 2) and suggest that robot gaze may be interpreted as a cue to its attentional state and infer referential intentions from this (supporting level 3).

Experiment 3 served to decide whether the effect of robot gaze on utterance comprehension is due to reflexive gaze-following inducing a purely "bottom-up" visual attention shift (level 2), or whether this is rather caused by the expectation that robot gaze reflects referential intentions such that people indeed establish joint attention with the robot (level 3, offering support for level 4 potentially observable in a different scenario). To shed light on these underlying processes, participants were asked to perform a correction task rather than simply validating utterances. A verbal correction of an utterance involving (especially incongruent) multimodal references implicitly requires participants to decide which referent they think the robot intended. Results indeed suggest that people not only reflexively follow this cue but consider the robot to indeed *see* and process visual information such that listeners assign an attentional and even intentional state to the robot, providing support for the application of response level 3.

Finally, the importance of human-like robot gaze and speech alignment is explored in Experiments 4 and 5. While Experiments 1-3 suggest that robot gaze affects utterance comprehension when it is aligned to robot speech in a manner that is typical for human gaze and speech production, it is an open question whether such alignment is necessary for people to interpret (robot) gaze as a an attentional/intentional cue. Thus, alignment was manipulated with respect to temporal synchronization and linear order of referential cues, investigating how gaze is interpreted and integrated into the incremental process of understanding an utterance. Results from these experiments illuminate the nature of robot gaze influence and, additionally, provide evidence supporting response level 3.

# 3. Gaze Following and Utterance Comprehension

In situated face-to-face communication human listeners are known to direct their visual attention based on both the speaker's utterance (Griffin and Bock, 2000) and speaker's gaze (Hanna and Brennan, 2007). In this chapter we examine, firstly, whether such behavior applies in HRI as well and, secondly, how robot speaker gaze influences utterance comprehension. That is, two experiments are presented that explore whether people respond to a robot showing referential gaze on level 1 (ignoring robot gaze), 2 (following robot gaze, possibly reflexively) and 3 (interpreting robot gaze as an attentional cue that may be used to establish joint attention) as described in Section 1.4.

Specifically, Experiment 1 examines whether people follow robot gaze and look towards an object fixated by the robot and, further, investigates whether robot gaze can help people to identify the referent of a referentially ambiguous utterance. Such a behavior could suggest that people interpret robot gaze similar to human gaze which, in turn, hints at an assignment of attentional and intentional states to the robot. Experiment 2 examines whether such referential robot gaze is in fact beneficial for the comprehension of referentially unambiguous robot utterances. The benefits of gaze were examined through manipulation of gaze congruency with respect to speech, simultaneously exploring consequences of a conflict between these referential cues.

Experiments 1 and 2 were run as a combined study, in order to simultaneously investigate very simple effects of robot gaze in Experiment 1 and more general effects of robot gaze in Experiment 2. For balancing conditions in items and fillers, we consider items of one study as additional fillers for the other study. Therefore we recorded data for both experiments from each participant in Experiment 1 and 2 and analyzed them separately as explained below. In both experiments, participants were generally required to pay close attention to the robot's utterance as well as the scene in order to quickly complete their task.

## 3.1. Experiment 1

One reason listeners may pay attention to speaker gaze is to identify the intended referent. Gaze may be particularly important when the utterance fails to uniquely identify an object in the shared environment. Consider this sample scenario: Person A and B are facing each other. A has her hands in cookie dough and needs B to pass her the next ingredient. There are two bowls, one to A's right and one to her left. The right one is filled with dark chocolate chips, the left bowl is filled with white chocolate chips. Thus, A says: "B, could you pass me this bowl please?" and briefly looks at the bowl to her right. Even though B might favor white chocolate chips, it is very likely that she will (rightly so) pass A the dark chocolate chips.

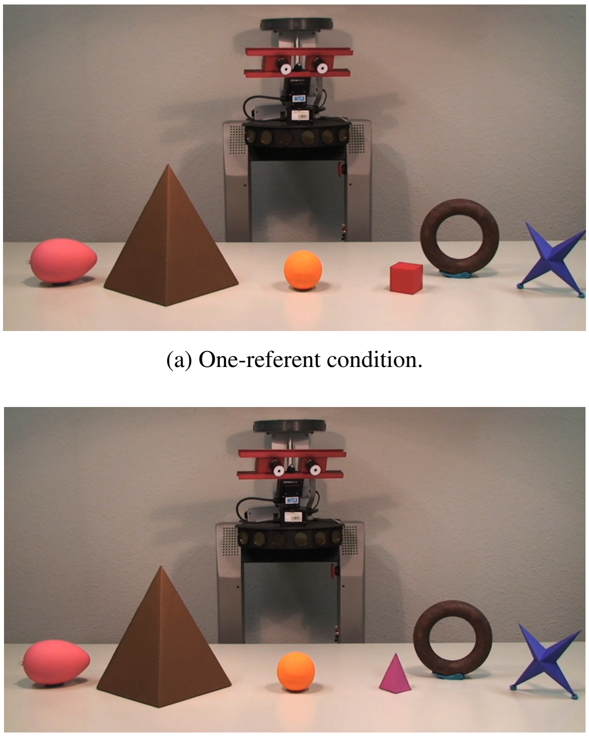
Obviously, people follow both the speaker's utterance as well as her gaze in order to understand the speaker's intention. The processes involved in understanding a spoken sentence have been studied extensively, as reported in Chapter 2. Previous findings suggest that people process a sentence incrementally with each new information further constraining the domain for interpretation: Listeners' eye movements indicated when and what people considered as potential referents of the sentence (Tanenhaus et al., 1995; Allopenna et al., 1998; Altmann and Kamide, 1999; Sedivy et al., 1999; Knoeferle et al., 2005). Moreover, people typically follow their partner's gaze and seek to establish shared attention in order to infer intentions and emotional states (Baron-Cohen et al., 1997a,b). Since speaker gaze is closely coupled to the utterance, it reveals what the speaker plans to mention (Meyer et al., 1998; Griffin and Bock, 2000). Thus, paying attention to these gaze cues and their referents can be beneficial for the listener – in particular, when the spoken utterance contains referential ambiguity, as illustrated in the example above.

All these findings together suggest that listeners' visual attention is influenced by both the spoken utterance and the speaker's gaze. Experiment 1 examines whether robot gaze is a similarly powerful cue which interlocutors follow and use in order to ground and possibly disambiguate robot utterances. Beyond the purely behavioral findings, linking robot gaze to the robot's utterance would indicate that people implicitly assign attentional states to the robot which connect the robot's visual perception with its spoken utterance. To begin investigating this hypothesis and, specifically to explore to what extent people retrieve referential information from robot gaze, we examined whether listeners followed robot gaze, both in cases when the utterance uniquely identifies the referent and, more interestingly, when there are several possible referents.

Thus, participants in this study faced a videotaped robot and were asked to judge its utterances for validity with respect to the shared scene.

Participants saw the robot while it produced a statement about several objects in its view and were asked to indicate whether or not the statement was valid. A description such as "The sphere is next to a cube." was accompanied by robot gaze movements – first to the sphere and then to the cube – each occurring shortly before the robot utters the corresponding noun phrases (Figure 3.1). To determine whether listeners followed robot gaze to a mentioned referent, we manipulated the referential ambiguity of the second noun phrase with respect to a given scene. That is, in one condition, the video showed among other shapes one sphere, one cube and one pyramid. In the second "two-referent" condition, there were two pyramids in the scene, both matching the referentially ambiguous utterance "The sphere is next to a pyramid". Consequently, we manipulated the single factor (Ambiguity) with two levels (one-referent, two-referent) within subjects.

Since participants were required to verify the statement against the scene, we assumed that their gaze behavior would be influenced by the robot's utterance. It was unclear, however, what impact robot gaze would have on participants' visual attention and their comprehension. In the one-referent condition, both robot gaze and speech identified a unique *target* object (a single cube). In the two-referent condition, the robot's utterance identified two potential referents (a *target* and a *competitor* pyramid) while robot gaze is directed only towards one pyramid (*target*). We observed and compared participants' looks towards the target and competitor objects in both conditions to establish whether people follow robot gaze. In the one-referent condition, we expected people to fixate the target at latest upon hearing it mentioned (Allopenna et al., 1998) and possibly earlier when people follow robot gaze. Crucially, in both conditions robot gaze was not required to determine the statement's validity since for either referent (target or competitor) the utterance was valid. In the two-referent condition, we expected people to fixate the target object upon mentioning if they generally follow gaze, or else that people would inspect both the target and competitor equally often. Moreover, if people considered the looked at pyramid as uniquely identified, response times for both conditions should be similar, suggesting that the linguistic ambiguity does not affect comprehension.



(a) One-referent condition.

(b) Two-referent condition.

(a)

Original sentence: ”Die Kugel ist neben einem Würfel.”

(Translation: ”The sphere is next to a cube.”)

(b)

Original sentence: ”Die Kugel ist neben einer Pyramide.”

(Translation: ”The sphere is next to a pyramide.”)

Figure 3.1.: Sample scenes from Experiment 1.

### 3.1.1. Method

#### Participants

Forty-eight native speakers of German, mainly students enrolled at Saarland University, took part in this study (34 females, 14 males). All participants reported normal or corrected-to-normal vision. Most of them had no experience with robots.

#### Materials

A set of 16 items was used in two conditions. In the one-referent condition, only one object in the scene matched the second noun phrase of the sentence. That is, all mentioned objects were uniquely identified by the uttered sentence. In the two-referent condition, two objects in the scene had the named shape. Thus, the ambiguity in the two-referent condition resulted from two potential referents in the scene as shown in Figure 3.1. There were equally many objects in each visual scene and the videos showed among other objects one sphere, one cube and one pyramid (in the one-referent condition). In the second, two-referent condition, there were two pyramids in the scene, both matching the utterance "The sphere is next to a pyramid." The result is a single factor with two levels that is manipulated within subjects. Crucially, the big brown pyramid (competitor object) was identical in both conditions such that looks to this object could be compared between conditions: When it was not a competitor for the referring noun "cube" in the one-referent condition compared to when it was a competitor object for the ambiguous referring noun "pyramid" in the two-referent condition. Since sentences were constructed to be valid for each potential referent (target, and competitor in the two-referent condition), items contained only true utterances. The task, however, was a decision task and we therefore created fillers that contained false statements (in total, 32 fillers were true and 24 were false) such that participants were required to pay attention to the utterances (57% true versus 43% false fillers).

We created 32 1920x1080 resolution video-clips showing the PeopleBot (Mobile Robots Inc., Amherst, NH, United States) robot onto which a pan-tilt unit was mounted, carrying the stereo camera. Note, that head orientation and eye-gaze of the robot are identical for this robot. The robot was positioned behind a table with a set of colored objects in front of it.

The objects are plain geometrical shapes of different colors and sizes. We used paper and styrofoam objects and colored them such that each object pair (of same shape) roughly had equally attractive colors in terms of saturation, e.g. red and orange, light green and grey or blue and green. In the one-referent condition (Figure 3.1a), each shape occurred only once on the table and the uttered sentence had a unique interpretation with respect to the scene. In the two-referent condition (Figure 3.1b), two objects of the same shape (but of different colors and sizes) were target and competitor referents in a corresponding sentence. The video-clips each showed a sequence of camera-movements (that are called *saccades* for the human eye) consecutively towards the object mentioned first and the target object mentioned second. Simultaneously, a synthesized sentence of the form given in Figure 3.1 was played back. Sentences were in German and synthesized using the Mary TTS system (Schroeder and Trouvain, 2001). Note, that the English determiner "a" has a unique interpretation while the original German determiner "ein(e)" may be understood as existential quantifier ("a") or as numeral ("one"). Thus, it is possible that incorrect, or rather undesired, responses in the two-referent condition would be elicited by a misinterpretation of the German determiner as numeral. However, we explicitly made participants aware of this issue and the accuracy rate for button presses suggests that participants did not use the numeral interpretation.

To align the synthesized sentences with the recorded scene, we first had to speed up the original video sequences by 140% so that the camera movements of the robot occurred at the appropriate point in the utterance. We subsequently overlaid the videos with the spoken stimulus sentences such that a robot fixation towards an object occurred one second prior to the onset of the referring noun, being consistent with corresponding findings on alignment of referential human gaze and speech production (Griffin and Bock, 2000; Van der Meulen et al., 2001). This also enabled us to observe two types of reactive human gaze: One being elicited by robot gaze (potentially indicating joint attention), the other being utterance-mediated shifts of visual attention (to inspect mentioned objects). In both conditions, participants had to give a positive answer since the statements were always true. Furthermore, across all 16 items, we balanced the stimuli with respect to target size (eight target objects are big and have small competitors and vice versa) and target location. For eight items, the target was placed to the left of the central object mentioned in the first noun phrase and in the other eight items it was placed to the right of the central object. Moreover, we have twelve different colors for twelve different object shapes that are employed as targets within our 16 items. Since a pilot study Staudte and Crocker (2008) suggested that participants initially inspected mainly the left area of the scene, we provided people with two seconds preview time which allowed them to inspect the entire scene before the robot started to move or speak.

In addition to the 16 items described above, we constructed 56 filler videos (of which 24 videos were experimental items for Experiment 2). Fillers contained between five and six objects and the location of the target object varied. Moreover, comparisons made in the robot's statements varied: In addition to location, color and size comparisons were used (e.g. "The heart is darker than the sphere", "The cone is shorter than the pyramid"). A complete list of item stimuli is provided in Appendix A (containing sentences) and B (containing still frames of each scene). We created two lists of stimuli, each containing 72 videos. Each participant saw only one version of an item and, in total, eight two-referent and eight one-referent items. The order of the filler videos was randomized for each participant individually such that an effect of trial sequence can be ruled out as explanation for the dependent variables' variance.

### Procedure

An EyeLink II head-mounted eye-tracker monitored participants' eye movements on a 24-inch monitor at a a temporal resolution of 500 Hz and a spatial resolution of 0.1 °. Participants were seated approximately 80 centimeters from the screen. Viewing was binocular, although only the dominant eye was tracked. The eye-tracker was adjusted, calibrated and validated manually for each participant using a nine-point fixation stimulus. Before the experiment, participants received written instructions about the experiment procedure and task: They were asked to attend to the presented videos and judge whether or not the robot's statement in each was valid with respect to the scene. In order to provide a cover story for this task, participants were told that the robot system would be evaluated. It still made many mistakes and participants' feedback was to be used as feedback in a machine learning procedure to improve the robot system. Crucially, gaze was typically not required nor did it change the assessment of sentence validity with respect to the scene (with an exception of only two fillers where sentence ambiguity affected validity). Each trial started with a fixation dot that appeared at the centre of the screen. Participants were instructed to always focus on that dot so as to allow the system to perform drift correction when necessary. Then a video was played until the participant pressed a button or until an overall duration of 12 seconds was reached. The entire experiment lasted approximately 30 minutes.

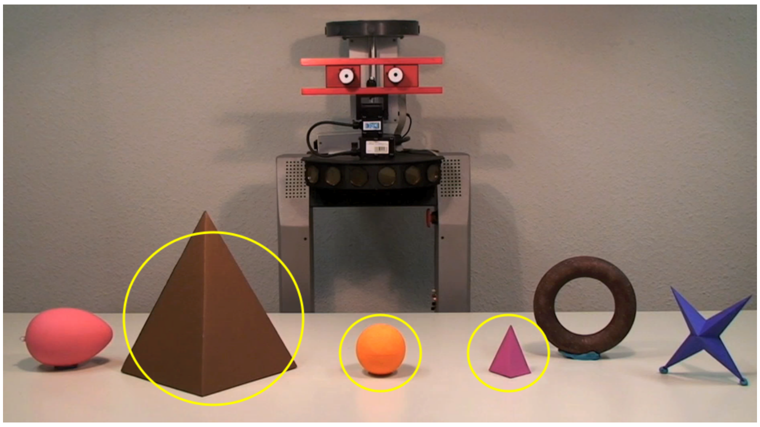


Figure 3.2.: Marked IAs in sample scene.

### Analysis

The presented videos were segmented into Interest Areas (IAs), i.e., each video contained regions that were labelled "anchor", "target" and "competitor" (Figure 3.2). The output of the eye-tracker was mapped onto these interest areas to compute the number of participant fixations on an object. The spoken utterance was a sentence as shown in Figure 3.1, describing the relation between a couple of objects. The noun "sphere" was encoded as the linguistic reference to the **anchor** object and the noun "cube"/"pyramid" was encoded as the linguistic **target** reference. In the one-referent condition, the referring noun "cube" uniquely identified one referent, namely the **target** object. The scene contained another **competitor** object also located next to the sphere. Since this object had a different shape (pyramid), however, it was not a possible referent for the second noun ("cube"). In the two-referent condition, in contrast, "pyramid" ambiguously identified the target *and* the competitor object since there were two pyramids in the scene.

The speech stream was segmented into two Interest Periods (IPs) based on the onsets and offsets of the encoded linguistic events. The IPs identify the time regions when the robot head fixated the target object and when it referred linguistically to the target object (see Figure 3.3). IP1 was defined as the 1000ms period preceding the onset of the target phrase, and contained the robot's fixation on the target object as well as some verbal content preceding the target noun phrase ("next to"). IP2 stretched from the target noun onset to offset and had a mean duration of 471ms (min=288ms, max=772ms). For

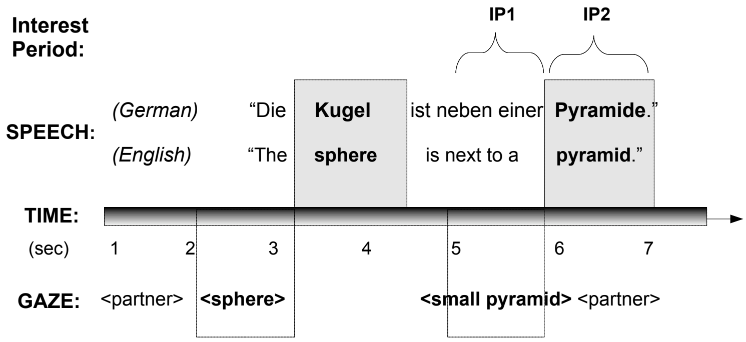


Figure 3.3.: The approximate timing of utterance-driven robot gaze for the given sentence.

the analysis of participants' fixations, all consecutive fixations within one IA and IP (i.e., before a saccade to another IA or the background occurred) were pooled and counted as one inspection. Trials that contained at least one beginning inspection towards an IA within an IP (coded as "1") are contrasted with trials that did not contain an inspection in the same slot ("0"). As a result, mean values represent inspection probabilities for a given IA/IP.

For the analysis of such un-accumulated, binary inspection data, in general, we used logistic regression (mixed-effects models with a logit link function from the *lme4* package in R Bates, 2005). Participants and items were included as random factors. To asses the contribution of a fixed factor or an interaction of two factors to explaining the variance of the dependent variable, we performed model reduction/simplification. That is, we used a $\chi^2$ comparison between the model including and excluding the factor as predictor and compare the log-likelihoods and AIC/BIC (Baayen et al., 2008; Jaeger, 2008).[1] For the comparison between levels of a factor we reported coefficients, standard errors (*SE*) and Wald's Z. For post-hoc comparisons among individual conditions in case of more than one predictor, we also used subsets of the data for each level of one predictor and fitted models with only the second predictor. P-Values, although shown in the tables, are potentially anti-conservative (Baayen et al., 2008) so we rather refer to coefficients being larger than two *SE*s for indicating significance or, additionally, generate p-Values using Markov chain Monte Carlo (MCMC) sampling when possible (see e.g. Kliegl et al. (2007, in press) or Knoeferle and Crocker (2009) for previous use of this

[1] For model reduction, models were fitted by ML whereas final models are fitted using REML (see Crawley, 2007, p.634ff)

method). Unfortunately, this sampling method is currently available only for linear mixed-effects models, and not for logistic (generalized linear) models which are used to fit binary data such as eye-movement data.

The response time was calculated as time elapsed from the offset of IP2, which marks the end of the sentence, until the moment of the button press. Inferential statistics for response times are conducted using linear mixed-effects models.

### Predictions

If robot gaze is not followed, we expect participants to solely rely on the robot's utterance and, thus, to fixate the competitor object more often in the two-referent condition than in the one-referent condition. That is, inspections on the referent (target in one-referent condition) would not be expected before IP2, i.e., when the referring noun is uttered. This behavior would support the hypothesis that people ignore robot gaze (response level 1 identified in Section 1.4) and, as a consequence, suggest that people do not consider this robot to share biological and cognitive mechanisms and do not recognize its "gaze" movements as an expression of directing visual attention. Not following and, crucially, not using the robot's gaze would also predict longer response times for the two-referent condition since the ambiguous referring expression has to be dealt with.

If, in contrast, participants follow gaze, we expect to observe looks towards the target even before it is being mentioned (that is, in IP1) since the robot's gaze preceded the target mention. Observing gaze-following would imply at least response level 2 described in Section 1.4. That is, either robot gaze is followed reflexively as is the case with other direction-inducing cues (level 2), or it is (further) assumed to reflect attentional states in which case the robot's gaze direction would also affect reference resolution (level 3). Specifically, if people interpret robot gaze as reflecting visual attention and, consequently, try to establish joint attention, they should further continue to attend to the target rather than the competitor when the referring noun is mentioned (IP2), even in the two-referent condition. This would indicate that participants jointly attended to the target object and inferred the communicative intention of the robot to mention this target object. Thus, response times would be expected to be equally long for both conditions since people could use robot gaze effectively to quickly resolve the referentially ambiguous phrase.

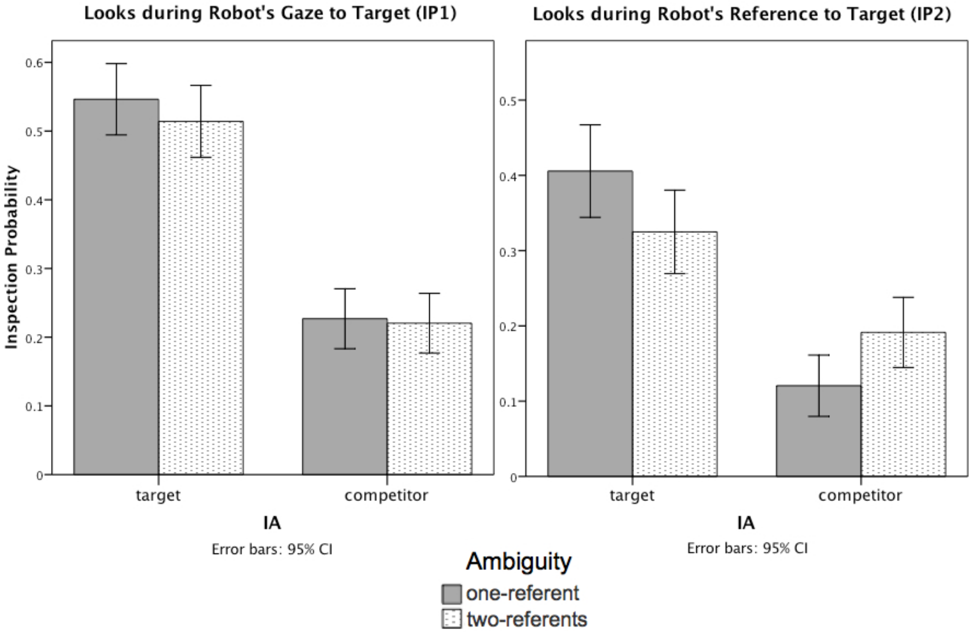


Figure 3.4.: Inspection proportions in both conditions per IA – for IP1 in the left graph and for IP2 in the right graph. IP2 shows a main effect of Condition as well as an interaction of Ambiguity and IA.

### 3.1.2. Results and Discussion

#### Eye movements

We observed that participants looked significantly more often at the target than at the competitor in both conditions and in both IPs. That is, Model1 reveals a main effect of factor IA[2] in IP1 ($\chi^2(1) = 152.24, p < 0.001$) as reported in Table 3.1. Separate analyses for each IA with respect to the factor Ambiguity were conducted by fitting Model2 to each IA and are also given in Table 3.1. The main effect of IA in IP1 indicates that participants did follow the robot's gaze towards the target object before it was even mentioned. Moreover, participants looked equally often at the target object in the two-referent and the one-referent condition (Figure 3.4), suggesting that participants followed robot gaze to the target even when there was another potential referent in the scene.

[2] It could be argued that IAs are not independent (more looks to one IA typically mean less looks to the other IA). However, in our counterbalanced design, looks to one object are considered target looks in some trials and

Table 3.1.: Fitted models on inspection data, Model1 includes IA as predictor, Model2 is fitted to separate data sets (IA=target/competitor), for both IP1 and IP2.

| IP1 | Predictor | Coefficient | *SE* | Wald Z | p |
|---|---|---|---|---|---|
| *(Model1)* | (Intercept) | −1.3731 | 0.1374 | −9.994 | <0.001 |
| | Ambig-one-referent | 0.0989 | 0.1183 | 0.836 | 0.40 |
| | IA-target | 1.4387 | 0.1195 | 12.033 | <0.001 |
| *IA = target:* | (Intercept) | 0.0374 | 0.2025 | 0.184 | 0.85 |
| *(Model2)* | one-referent | 0.1705 | 0.1613 | 1.057 | 0.29 |
| *IA = competitor:* | (Intercept) | −1.4223 | 0.2065 | −6.888 | <0.001 |
| *(Model2)* | one-referent | 0.0246 | 0.1867 | 0.132 | 0.89 |
| IP2 | | | | | |
| *(Model1)* | (Intercept) | -1.5108 | 0.1788 | -8.447 | <0.001 |
| | Ambig-one-referent | -0.5625 | 0.2526 | -2.227 | <0.05 |
| | IA-target | 0.7403 | 0.2021 | 3.662 | <0.001 |
| | one-referent:target | 0.9281 | 0.3116 | 2.978 | <0.01 |
| *IA = target:* | (Intercept) | −0.7609 | 0.1767 | −4.307 | <0.001 |
| *(Model2)* | one-referent | 0.3640 | 0.1865 | 1.952 | 0.051 |
| *IA = competitor:* | (Intercept) | −1.8074 | 0.2678 | −6.738 | <0.001 |
| *(Model2)* | one-referent | −0.6809 | 0.2702 | −2.520 | <0.05 |

$Model1 : Inspected \sim IA * / + Ambiguity + (1|subject) + (1|item),$
$family = binomial(link = "logit")$
$Model2 : IA \sim Ambiguity + (1|subject) + (1|item), family =$
$binomial(link = "logit")$

There was no main effect of Ambiguity in IP2. However, we found an interaction of Ambiguity and IA so we kept both predictors. The interaction can be interpreted such that, upon hearing the referring noun mentioned, participants looked more often towards the competitor object in the two-referent condition than in the one-referent condition. This increase in looks towards the competitor suggests that participants did notice the referential ambiguity. Nevertheless, there is a strong preference for fixating the target object in both conditions which indicates that participants identified the target by means of robot gaze despite the referential ambiguity.

### Response Times

Trials were removed when participants had pressed the wrong button (2%). We further excluded trials as outliers when the response time was $\pm 2.5 \times SE$ above or below a subject's mean (2.79 %). As predicted, we observed no significant difference in the response times (1438.4ms in the one-referent versus 1467.9ms in the two-referent condition). Fitting a mixed-effects model with Ambiguity as predictor ($RT \sim Ambiguity + (1|subject) + (1|item)$) shows that Ambiguity does not contribute to explaining the variance in the dependent variable Response Time ($\chi^2(1) = 0.0066, p = 0.935$).

The findings on both response time and the recorded eye movement data consistently suggest that people (a) follow robot gaze, and (b) use robot gaze to constrain the domain of interpretation, effectively resolving referential ambiguity. There are, however, several limitations to this study. Firstly, sentences in this experiment were referentially ambiguous which possibly emphasized the role of robot gaze even though the task did not require the use of gaze. That is, in the absence of sufficient linguistic information, any additional cue inducing a preference for interpretation (such as gaze, but also simple visual highlighting) may have similarly been used. Secondly, the absence of a response time effect is weak and indirect evidence for the facilitation of reference resolution. Thus, this initial study has provided convincing evidence for gaze-following behavior in HRI but the actual benefit of robot gaze cues for utterance comprehension requires further investigation.

Experiment 2 sought to examine the influence or robot gaze when accompanying unambiguous (one-referent) sentences compared against a baseline with neutral gaze, and compared to gaze that is directed at an irrelevant object.

competitor looks in other trials. Therefore target and competitor looks are to some extent independent. At this stage, we include analyses including and excluding IA as a factor although emphasis is put on separate analyses for each IA in the remainder of this thesis.

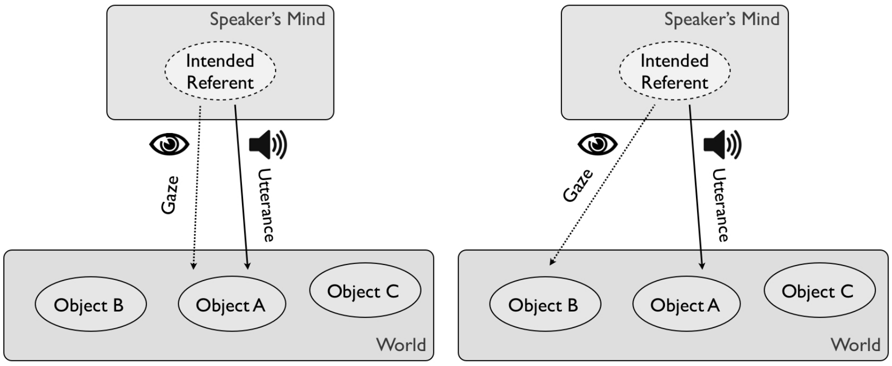
## 3.2. Experiment 2

Experiment 1 demonstrated that people follow robot gaze to an object prior to its mention and continue to fixate the object even when multiple objects are compatible with the spoken reference. However, the referential ambiguity of the statements may have enhanced the utility of an additional cue like gaze for reference resolution. Experiment 2 sought to confirm this result in the context of globally unambiguous sentences. That is, we examine whether referential robot gaze is followed and used for reference resolution even though utterances can be validated without paying any attention to robot gaze. Specifically, we investigate the actual benefit of robot gaze when accompanying sentences that contain only temporary referential ambiguity. The benefit will be assessed by comparison of response times for such a sentence when accompanied by referential robot gaze and when accompanied by neutral robot gaze. Observing such a beneficial effect of referential robot gaze would further suggest that people indeed establish a link between gazed at and mentioned objects via the assignment of attentional states to the robot.

When considering referential gaze and its utility for reference resolution, the question naturally arises whether referential gaze can also disrupt reference resolution if, for instance, it identifies an entity other than the one referenced in the utterance. Consider another example scenario: Person A is preparing cup-cake dough and needs person B to pass her yet another ingredient. There are two more bowls on the table, one filled with raspberries to her right and another one filled with blueberries to her left. Since A's hands are covered with dough, she says: "B, could you please pass me the *raspberries*?" If she looks at the raspberries already before mentioning them, B can use this early indicator to quickly identify A's referential intention, as shown by Hanna and Brennan (2007). If A looked at the *blueberries* instead of the raspberries, B is most likely confused about what she should do. Did A intend to say "blueberries" but was thoughtless and mentioned the wrong kind of berries? Or did A say what she meant but looked at the wrong bowl – maybe because she incorrectly remembered where she had put the raspberry bowl?

If people assign attentional states to the robot such that people would similarly assume that an object looked at by the robot is probably the one it intended to mention, then incongruent referential gaze may be an irritating cue that somehow disrupts utterance comprehension.

Thus, to further investigate the benefit of robot gaze, we manipulated the *congruency*



(a) Congruent multi-modal reference to one object. A was apparently intended by the speaker.

(b) Incongruent multi-modal reference to two different objects. Was A or B intended by the speaker?

Figure 3.5.: Multimodal references.

of our robot's gaze as a potential cue for intended meaning as well as the *validity* of the statements. Statements were either true or false, that is, the stated relationship between objects held or not, and the visual reference (established by robot gaze) was either congruent, incongruent or neutral with respect to the linguistic reference. We consider gaze to be congruent (and helpful) when it is directed towards the same object that is going to be mentioned shortly afterwards (reference match, see also Figure 3.5a) while it is considered as incongruent when gaze is directed to an object different from the mentioned referent (mismatch, Figure 3.5b). In a third congruency level robot, gaze was neutral. The robot briefly looked down at the scene and back towards the listener before beginning to utter a scene description. The neutral gaze behavior provided a baseline condition in which participants' visual attention was purely a response to the produced robot utterance and comprehension was uninformed by any joint attention mechanisms.

Robot statements were of the form given in the example sentence below. The second noun phrase was temporarily referentially ambiguous, providing time for participants to integrate the gaze cue with the ambiguous noun before the mentioned color disambiguated the referent.

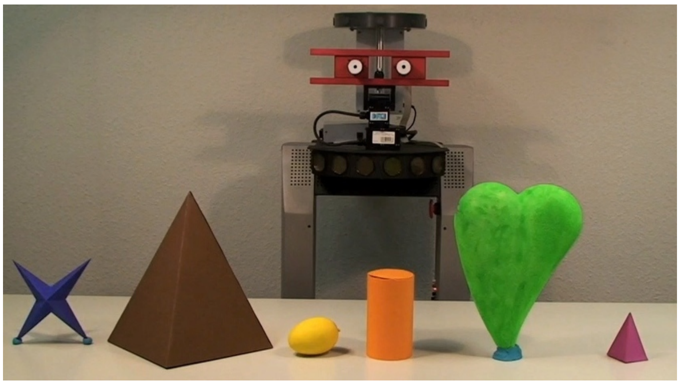


Figure 3.6.: Sample scene from Experiment 2.

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

The scene provided two potential referents (e.g., two pyramids of different sizes and colors) one of which the robot mentioned. One pyramid matched the description of the scene (*is shorter* than the cylinder) while the other did not (it is actually *taller* than the cylinder). Thus, which pyramid was finally mentioned depended on the color adjective and determined whether the statement was valid or not. The manipulation of both factors, Statement Validity and Gaze Congruency, resulted in six conditions per item. In Table 3.2, we provide an example for all conditions that the example sentence could appear in (given a corresponding scene depicted in Figure 3.6).

### 3.2.1. Method

#### Participants & Procedure

The same group of participants as in Experiment 1 was tested in an identical procedure.

#### Materials

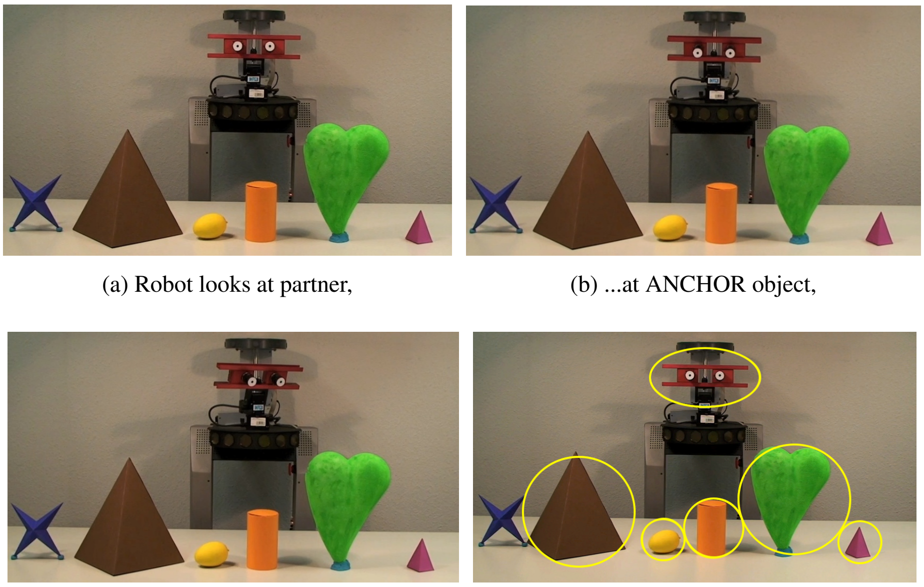
A set of 24 items was used. Each item consisted of three different videos and two different sentences, i.e., appears in six conditions as shown in Table 3.2. Additionally we counterbalanced each item by reversing the comparative adjective, for instance, from

Table 3.2.: Given the scene in Figure 3.6, manipulation of sentence validity and robot gaze results in these six conditions.

| Robot Gaze | **True** Sentence: *"The cylinder is taller than the pyramid that is pink."* | **False** Sentence: *"The cylinder is taller than the pyramid that is brown."* |
|---|---|---|
| **Congruent** | looks to small pink pyramid | looks to big brown pyramid |
| **Incongruent** | looks to big brown pyramid | looks to small pink pyramid |
| **Neutral** | – | – |

″taller″ to ″shorter″, such that the target becomes the competitor and vice versa. We obtained a total of twelve videos per item while ensuring that target size, location and color were balanced. All versions showed the same scene and only differed with respect to where the robot looks and whether it verbally refers to the correct (target) object. Twelve different object shapes appeared twice each as target-competitor pairs to produce 24 items. For each shape, we created three different sizes (small, medium, large) and used each small-large pair as target-competitor pairs and the medium sized shape as anchor for another target-competitor pair. Moreover, each scene contained three additional distractors, two were large and small and positioned to either side of the anchor. They served as potential competitors for partial utterances up to the comparative (e.g. "The pyramid is taller than"). The third distractor was typically small and positioned to the far left or far right of the scene.

Prior to the experiment, target-competitor pairs were pre-tested in order to make sure that their size and color differences were easily recognizable. We used a questionnaire that showed photographs of the original scenes excluding the robot. Twenty participants judged whether a given item sentence accurately described what was visible in the scene. For each scene, three sentences were given and only one of those contained a comparison between item objects (anchor and target/competitor). Overall, 50 % of the sentences were true and 50 % were false in order to avoid an acquiescence bias. A 7-level Likert scale from 1 (incorrect) to 7 (correct) allowed for a graded judgement of the sentences′ validity. The results exhibit a mean deviation of 0.26 points from the op-



(a) Robot looks at partner,

(b) ...at ANCHOR object,

(c) ...at TARGET object,

(d) ...and back up. (with marked regions of interest)

Figure 3.7.: Sequence of gaze movements in sample scene from Experiment 2.

timal answer (1 and 7) and no outlier items which clearly shows that the comparisons between the distinct objects and their sizes are clear and easily assessable.

We created 24 items each in 12 versions (6 conditions, each counterbalanced), obtaining a total of 288 item videos of the same type that we did for Experiment 1. The robot fixations and the spoken sentence were again aligned such that a fixation towards an object occurred one second prior to the onset of the referring noun.

Moreover, we constructed 48 filler videos (16 videos that were item trials for Experiment 1 and an additional set of 32 filler videos) such that we obtained twice as many fillers as we had items. Half of the experimental items were correct, i.e., in a *true* condition, and one third was true and showed congruent or neutral gaze. To compensate for the relatively high proportion of anomalous items, a large number of fillers contained a correct statement and congruent robot gaze behavior. That is, 36 of 48 filler videos in total contained true statements (75%) and 24 were both true and congruent (50%). This results in an overall distribution of 66% true trials in the experiment. This bias towards true statements was intended to maintain the participant's trust in the competence of the robot. However, robot gaze can be considered relatively unpredictive since there were only 40 congruent trials overall (55.5%) showing robot gaze to an object that was subsequently mentioned. This reduces the likelihood of gaze-following emerging for purely strategic reasons.

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 randomized for each participant individually with the constraint that between items at least one filler was shown.

### Analysis

The Interest Areas (IAs) in this experiment consisted of the anchor, the target and the competitor objects, the robot head and the two distractors next to the anchor (see Figure 3.7d). The temporarily ambiguous target noun "pyramid" from the example sentence above was the *spoken reference* to two potential objects (*referents*) in the scene: the small pink **target** pyramid or the large brown **competitor** pyramid, and referential robot gaze provided a *visual reference* to one of these objects. The small pink pyramid was considered as *target object* because the partial description "The cylinder is *taller than* the pyramid" applied to the small pink pyramid. That is, the sentence-final mention of the adjective "pink" resulted in a *correct* statement whereas mentioning the brown pyramid resulted in an *incorrect* comparison.

We segmented the speech stream into three Interest Periods (IP) as depicted in Figure 3.8. IP1 was defined as the 1000ms period ending at the onset of the target noun "pyramid". It contained the robot's fixation towards the target object as well as some verbal content preceding the target noun. IP2 stretched from target noun onset to offset. It had the same mean duration of 471ms as in Experiment 1 which was constant for all conditions of an item. IP3 was defined as the 700ms period beginning at the onset of the disambiguating color adjective.

This adjective denoting the color of the referent completed the linguistic reference and unambiguously identified the actual referent (IP3). Only at that point in time was it possible to judge the statement validity, which is why it is called the linguistic point of disambiguation (LPoD).[3] The elapsed time between this adjective onset and the moment of the button press was therefore considered as the response time.

[3] A similar design, also featuring late linguistic disambiguation with early visual disambiguation by means of gaze-following, was presented in an HHI scenario by Hanna and Brennan (2007) .

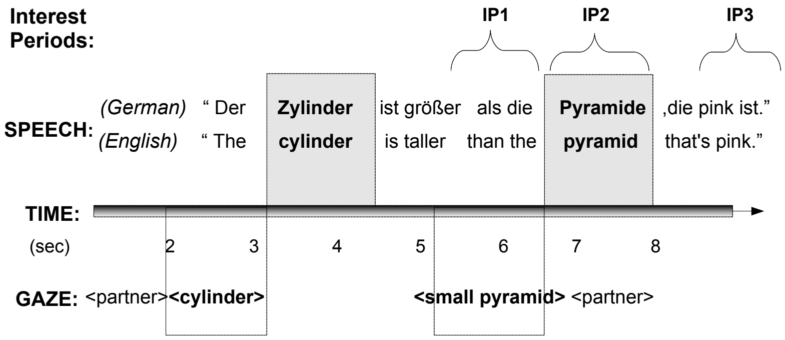

Interest Periods:
IP1
IP2
IP3
SPEECH:
(German) " Der Zylinder ist größer als die Pyramide ,die pink ist."
(English) " The cylinder is taller than the pyramid that's pink."
TIME:
(sec)
2
3
4
5
6
7
8
GAZE: <partner> <cylinder> <small pyramid> <partner>

Figure 3.8.: The approximate timing of utterance-driven robot gaze, in a true-congruent condition.

Mixed-effects models (lmer) were used to fit both eye-movement and response time data. Participants and items were included as random factors, and Gaze Congruency (as well as Sentence Validity in IP3) were included as fixed factors.

### Predictions

Since participants had to validate the utterance with respect to a given scene, we expected participants' gaze to be mediated by robot speech. That is, we predicted that during sentence processing people would look at entities according to the incrementally constrained set of possible referents (Tanenhaus et al., 1995; Sedivy et al., 1999). Since the second referent was not uniquely identified until the end of a sentence (LPoD), participants could keep several hypotheses about potential referents until then. We expected listeners' gaze throughout a trial to indicate which hypotheses about referent(s) the listener currently maintained.

Based on the findings of Experiment 1, we hypothesized that people may follow not only robot speech but also robot gaze (as explained in Section 1.4). In this experiment, we therefore expected in those conditions showing referential gaze (congruent and incongruent) to similarly observe gaze-following. In particular in IP1, when robot gaze was directed towards either the target or the competitor while none of them was yet linguistically identified, fixations are expected to reveal whether gaze-following occurred or not.

In Section 1.4, we further differentiated between different mechanisms involved in gaze-following: People might reflexively orient towards the cued direction (response level 2) and/or people might follow robot gaze because they consider the robot's gaze movements to reveal information about the robot's visual attention states (response level 3). Cueing effects resulting from reflexive gaze-following are relatively short-lived (Friesen and Kingstone, 1998; Driver et al., 1999; Langton and Bruce, 1999) and should have disappeared by the end of the sentence. However, it is conceivable that information obtained through an involuntary attention shift affects further processing. Yet, we would expect integration of such information during utterance comprehension only if people consider it relevant and, thus, use it voluntarily – but this remains an open issue at this stage. In contrast, if people follow robot gaze in order to attend to the same object as the robot, that is, to establish joint attention with it as is the case in human-human interaction (Hanna and Brennan, 2007), we would clearly expect robot gaze to affect both people's fixation behavior, even beyond IP1, as well as response times.

In IP2 we, thus, expected a continued preference to inspect the object previously identified by robot gaze. In the neutral gaze condition, inspections might reveal whether people use the partial utterance to constraint the domain of interpretation. That is, the target would be inspected more frequently than the competitor since only the target is consistent with the utterance so far ("*The cylinder is taller than the pyramid*").

Since IP3 revealed the match (congruent condition) or mismatch (incongruent condition) of visual and linguistic references, we predicted that a match would cause people to continue to inspect the object they were already looking at (presumable the object identified by robot gaze, if listeners followed robot gaze). A mismatch in referential cues would, thus, lead to an attention shift from the visual referent – the target in condition false-incongruent and the competitor in condition true-incongruent – to the object identified by the color adjective – the competitor in condition false-incongruent, or target in condition true-incongruent. Thus, in the congruent condition, we predicted inspections mainly on the consistently identified referent whereas inspections on both the visual and the linguistic referent were expected in the incongruent condition.

Furthermore, we would also expect a main effect of Gaze Congruency for response times: If participants exploit robot gaze and assume that it indicates the robot's focus of visual attention, they will correctly anticipate the validity of statements when gaze is congruent. In contrast, when gaze is incongruent with the statement, we would predict that participants anticipate a proposition that eventually does not match with the actual robot statement. Hence, slower response time for incongruent robot gaze would be expected. Since neutral gaze neither facilitated nor disrupted the judgement of the statement validity, we predicted intermediate response time for this condition.

Crucially, if people followed and used robot gaze for purely strategic reasons, their behavior should change after a few trials when people realize that robot gaze is almost equally often misleading as it is helping to anticipate the correct referent. Furthermore, as true statements are more frequent and expected to elicit faster response times than false statements, a main effect of statement validity was expected for response times.

### 3.2.2. Results and Discussion

#### Eye movements

Figures 3.9 and 3.10 show plots of the eye-movement data for the whole duration of a trial and for each condition individually. The initial two seconds of a trial were preview time, the robot head started moving approximately 2,000ms after trial start. Plotting begins after preview time and ends at 10,000ms, just after the end of the robot utterance. This 8,000ms-window is divided into 250ms-bins and fixation proportions are computed for each IA (anchor, target, competitor and robot head) within each bin. Fixations that did not fall within an IA were counted towards background fixations and are not included in the graph. The average onset of IP1 is at 5,913ms after plot beginning, the average onset of IP2 is consequently at 6,913ms with an average duration of 471ms. The average onset of IP3 is at 7,949ms and lasts, by definition, 700ms.

Each plot in the time graphs shows that people initially looked mainly at the robot head. When the robot head moved towards the anchor and, more clearly, when the robot started speaking, people directed visual attention away from the robot's head and towards the anchor. Throughout the course of a trial people rarely looked back at the robot head. The plots, however, clearly indicate gaze-following which suggests that people used robot gaze peripherally. Gaze-following is indicated by people's looks in IP1 towards either the target (in true-congruent and false-incongruent conditions) or to the competitor (true-incongruent, false-congruent), following the robot's gaze towards these objects. Consequently, in conditions true-neutral and false-neutral, neither target nor competitor were being closely attended in IP1. Notably, in the presence of robot gaze people started fixating either target or competitor even before IP1 began. This is most likely due to the long time window that a robot 'saccade' spans. Since IP1 began with the end of the camera movement towards an object such that the robot fixated this object approximately one second prior to noun onset, the actual movement (or saccade) towards the object preceded IP1. Plots of congruent conditions show that people more frequently fixated the looked at and mentioned object until the end of the trial

Table 3.3.: Models fitted to separate inspection data sets (IA=target/competitor), in IP1 and IP2.

| | Predictor | Coefficient | *SE* | Wald Z | p |
|---|---|---|---|---|---|
| IP1 | | | | | |
| *IA = target* | (Intercept) | -2.1457 | 0.1901 | -11.290 | <0.001 |
| | competitorgaze | -0.6764 | 0.2797 | -2.418 | <0.05 |
| | targetgaze | 1.6323 | 0.2032 | 8.032 | <0.001 |
| *IA = competitor* | (Intercept) | -2.4090 | 0.1994 | -12.081 | <0.001 |
| | competitorgaze | 1.6547 | 0.2222 | 7.448 | <0.001 |
| | targetgaze | -0.5496 | 0.3070 | -1.790 | 0.073 |
| IP2 | | | | | |
| *IA = target* | (Intercept) | -1.0967 | 0.1825 | -6.009 | <0.001 |
| | competitorgaze | -1.2799 | 0.2366 | -5.409 | <0.001 |
| | targetgaze | 0.4785 | 0.1857 | 2.577 | <0.01 |
| *IA = competitor* | (Intercept) | -2.1023 | 0.2012 | -10.450 | <0.001 |
| | competitorgaze | 1.5116 | 0.2145 | 7.048 | <0.001 |
| | targetgaze | -0.3726 | 0.2847 | -1.309 | 0.191 |

$Model : IA \sim Gaze + (1|subject) + (1|item),$
$family = binomial(link = "logit")$

while paying little attention to the other, potentially competing object. In incongruent conditions, people mostly fixated the looked at object in IP1 and IP2 (where the referring expression is still ambiguous) and then fixated the object identified by the color adjective in IP3.

Since sentence truth did not play a role in IP1 and IP2 (because the LPoD only occurs in IP3), we collapsed each two conditions where trials were identical up to IP2 for further inspection analyses. That is, conditions true-congruent and false-incongruent were collapsed into the condition "gaze to target", true-incongruent and false-congruent were collapsed into the condition "gaze to competitor" and the two neutral conditions

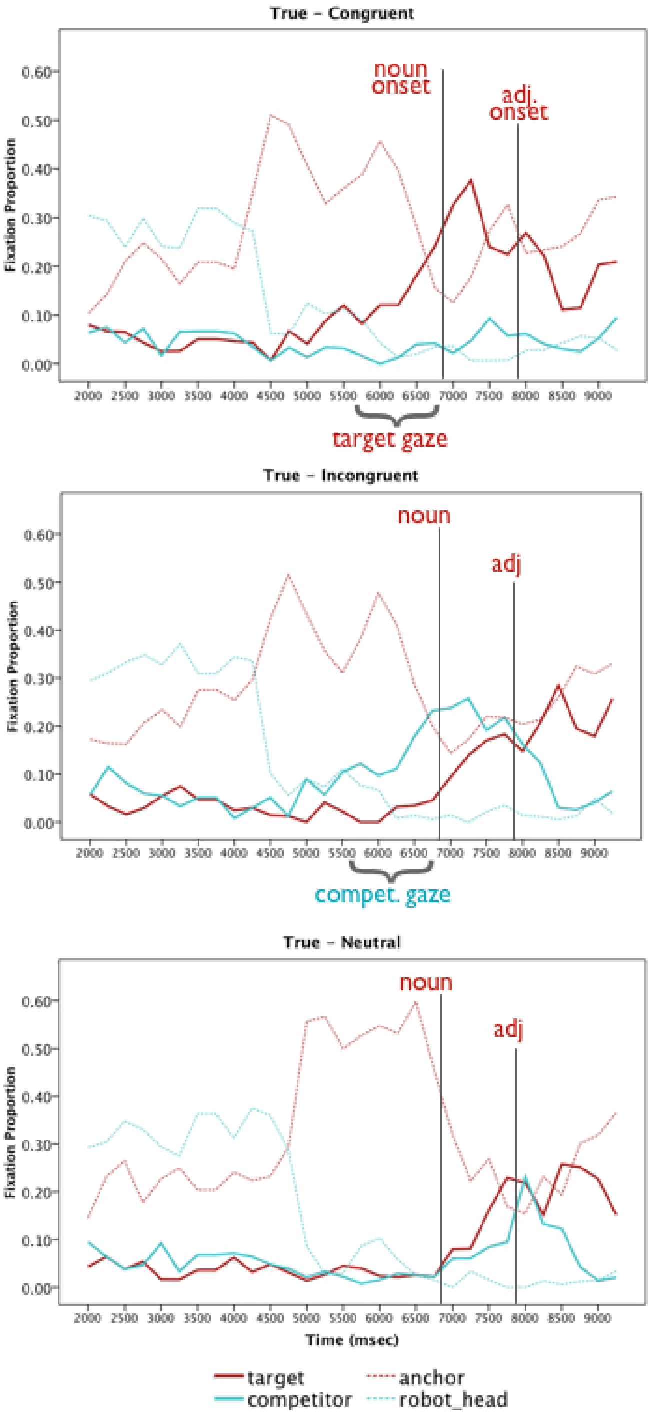


Figure 3.9.: Average Fixation proportions calculated using 250ms bins, for true sentences. IP1 ends on noun onset and contains robot's gaze towards target/competitor, IP2 stretches from the (ambiguous) noun onset to offset, and IP3 comprises the disambiguating color adjective.

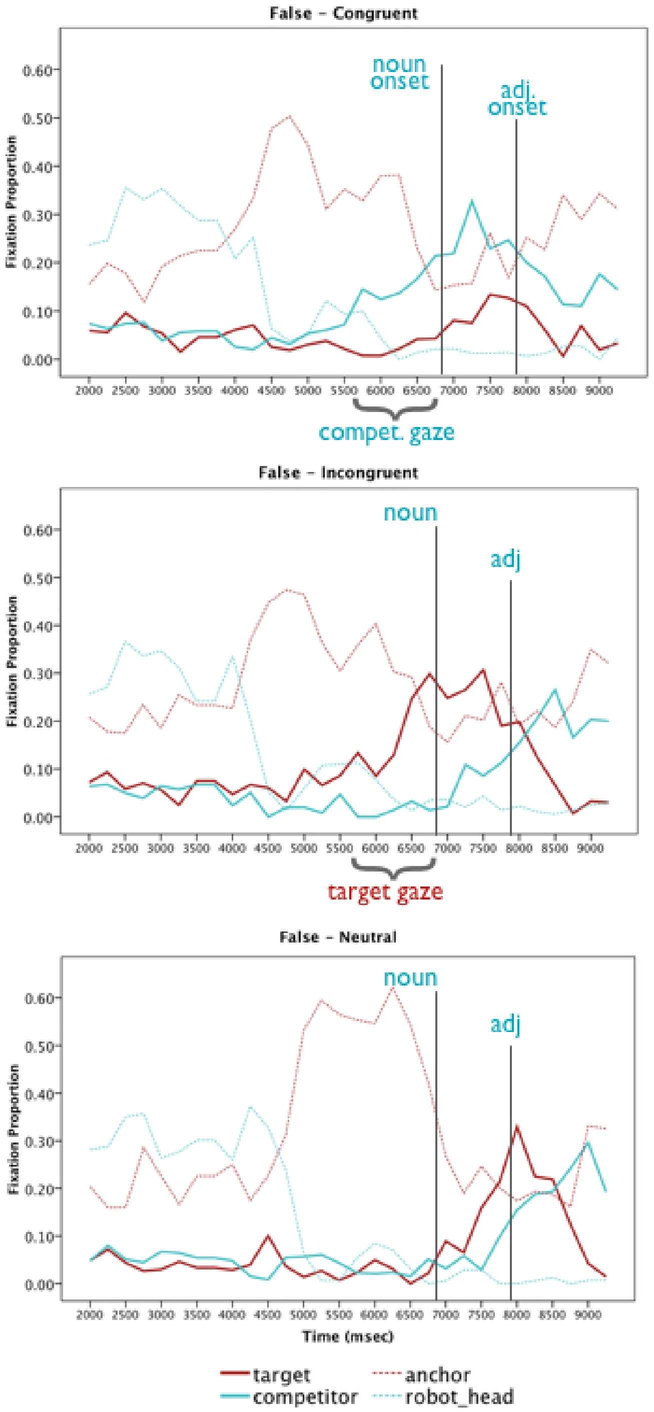


Figure 3.10.: Average Fixation proportions calculated using 250ms bins, for false sentences. IP1 ends on noun onset and contains robot's gaze towards target/competitor, IP2 stretches from the (ambiguous) noun onset to offset, and IP3 comprises the disambiguating color adjective.

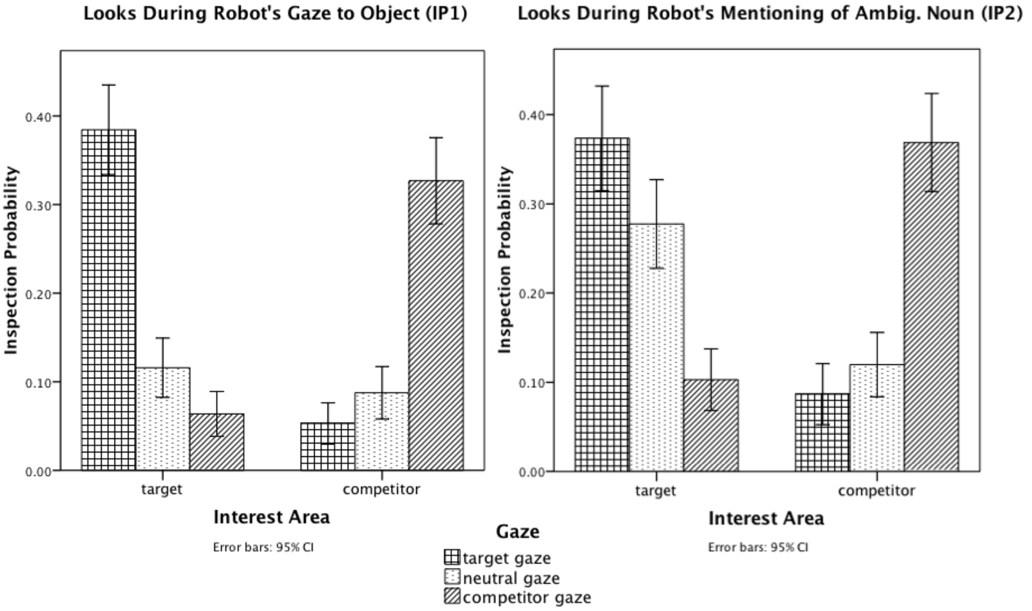


Figure 3.11.: Mean inspection probabilities in three gaze conditions for IP1 (left graph) and IP2 (right graph). IP1 is the 1,000ms time window preceding the target noun onset. IP2 stretches from target noun onset to offset.

were merged to one "neutral"-condition.

Results from inferential statistics for IP1 and IP2 are given in Table 3.3. Summarizing these numbers, we observed the following:

*IP1*("The cylinder is TALLER THAN THE"): During IP1, robot gaze is the only potential cue to the intended target (e.g. big brown or small pink pyramid). In this IP, robot gaze had a main effect on people's inspection behavior (visible on the target IA: $\chi^2(2) = 138.97; p < 0.001$ and also the competitor IA: $\chi^2(2) = 118.17; p < 0.001$). The graph in Figure 3.11 depicts people's inspections on the target and competitor IAs and shows that people inspected the target IA with a significantly higher probability when the robot also looked at the target than when it looked at the competitor or showed neutral gaze. Similarly, when the robot looked at the competitor, we observed significantly more inspections on the competitor than in the other conditions. In contrast, when robot gaze was neutral, inspections to both IAs were equally unlikely at this point. According to previous work on sentence processing (Eberhard et al., 1995; Sedivy et al., 1999), the mentioned comparative could in fact constrain the domain of interpretation already at this point such that the target becomes a more likely referent than the competitor. However, this preference was not yet visible in fixations patterns (but will be in IP2). One reason for this may be that target and competitor are on the far sides of the table and not so salient at this point, another reason might be that there are other objects in the scene which also match the utterance so far (at least one of the distractor as discussed below).

*IP2*("The cylinder is taller than the PYRAMID"): The inspection pattern observed in IP1 persisted in IP2 for both conditions showing referential robot gaze (target IA: $\chi^2(2) = 64.8; p < 0.001$ and competitor IA: $\chi^2(2) = 87.53; p < 0.001$). For neutral robot gaze, participants were now more likely to inspect the target IA (small, pink pyramid) than the competitor which was consistent with the incomplete utterance so far. Pairwise comparisons between target and competitor inspections for neutral gaze in IP2 showed that people inspected the target rather than the competitor ($p < 0.001$). Based on the comparative in the sentence (taller/shorter), the target was the more probable referent. However, referential robot gaze introduced additional (and potentially conflicting) information since it drew attention to either target or competitor prior to IP2. Thus, when referential robot gaze was available, the preference for the object that met the linguistic constraints of the utterance (target) was no longer observable. Interestingly, participants simply followed the robot's gaze to either the target or the competitor instead.

For IP1, we conducted a similar analysis (as for target and competitor) for both distractor objects next to the anchor (see Figure 3.6). Since there was always one tall and one short distractor, one distractor always matched the linguistic constraints in IP1 (was shorter/taller that the cylinder) while the other did not. Moreover, a short distractor was between the anchor and the tall target/competitor while a tall distractor was located between anchor and short target/competitor object. Thus, when the robot uttered the correct sentence "The cylinder is taller than the (pyramid that is pink)" and gazed incongruently at the tall, brown pyramid, its continuous gaze movement would pass the short distractor. This implies that, in addition to the linguistic constraints (the comparative), robot gaze potentially also constrained or biased interpretation.

Since the size of the distractor did not affect whether the sentence comparative induced a preference for it or not (the same main effect was observed for both distractor objects, short and tall, separately), we collapsed the two IAs for this analysis and obtained the following factors: Sentence Comparative Match (distractor size either matched or mismatched the comparative in the sentence) and Gaze Direction Match. That is, gaze was either neutral, or the distractor was in the general direction of robot

Table 3.4.: Model fitted to inspection data on distractor IAs, for IP1.

| Predictor | Coefficient | *SE* | Wald Z | p |
|---|---|---|---|---|
| (Intercept) | 0.3858 | 0.1022 | 3.777 | <0.001 |
| Sentence- mismatch | -0.6168 | 0.0949 | -6.499 | <0.001 |
| Gaze - mismatch | -1.6094 | 0.1236 | -13.016 | <0.001 |
| Gaze - none | -0.3237 | 0.1083 | -2.989 | <0.005 |

$$Model : Inspected \sim SentenceMatch + GazeDirection$$
$$+(1|subject) + (1|item), family = binomial(link = "logit")$$

gaze (i.e., when the robot looked at the target/competitor located further away its gaze passed this distractor), or the distractor was in the opposite direction to robot gaze.

The results shown in Figure 3.12 suggest that the direction of the robot's gaze was indeed a very dominant cue that mainly determined where people looked. Nevertheless, we found main effects for both robot Gaze Direction Match ($\chi^2(2) = 201.94; p < 0.001$) and the Sentence Comparative Match ($\chi^2(1) = 42.44; p < 0.001$, see also Table 3.4). That is, even though people's visual attention was primarily influenced by robot gaze, subtle linguistic information (such as the comparative) was also picked up and incrementally constrained the domain of interpretation, as reflected by inspection probabilities. This shows that people pay close attention to both modalities, speech and gaze.

*IP3*("The cylinder is taller than the pyramid that is PINK/BROWN."): IP3 contains the linguistic point of disambiguation (LPoD) specifying which pyramid is indeed being mentioned. This IP was considered separately from IP1 and IP2 since both factors, Sentence Validity and Gaze Congruency, now affected participant behavior. Fitting and comparing various linear mixed-effects models (for both IAs separately) showed that both predictors and, more importantly, their interaction significantly contribute to a model of the respective data set. That is, model reduction (and model parameters in Table 3.5) reveals a robust main effect of Statement Validity ($\chi^2(1) = 24.47; p < 0.001$), that is for both IAs, the validity level "true" is significantly different from the intercept level ("false"). The positive coefficient of the predictor level ('true') indicates a higher inspection probability for the given IA compared to the intercept level ('false'). This

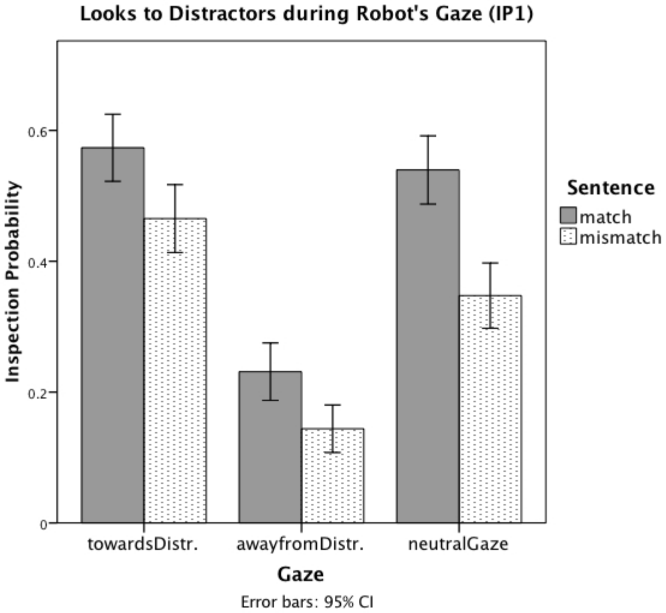


Figure 3.12.: Inspection proportions on distractor object in two gaze conditions (towards/away from distractor) for both comparatives (match/mismatch with distractor size).

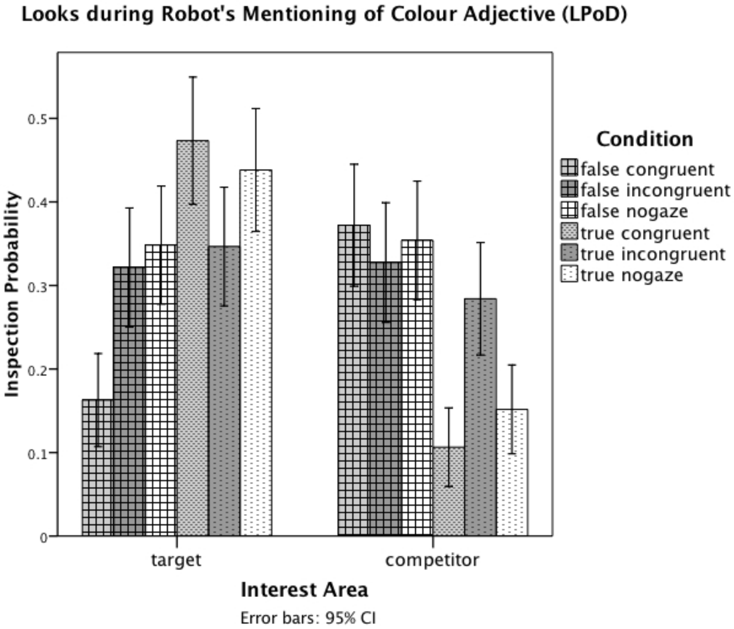


Figure 3.13.: Looks to target/competitor during adjective-mentioning (IP3), in each condition.

Table 3.5.: Models fitted to separate inspection data sets (IA=target/competitor), in IP3.

| | Predictor | Coefficient | *SE* | Wald Z | p |
|---|---|---|---|---|---|
| *IA = target* | (Intercept) | -1.7067 | 0.2177 | -7.840 | <0.001 |
| | Validity - true | 1.5996 | 0.2611 | 6.128 | <0.001 |
| | Congr - incongruent | 0.9209 | 0.2679 | 3.437 | <0.001 |
| | Congr - neutral | 1.0457 | 0.2630 | 3.977 | <0.001 |
| | true:incongruent | -1.4702 | 0.3488 | -4.215 | <0.001 |
| | true:neutral | -1.1999 | 0.3414 | -3.514 | <0.001 |
| *IA = competitor* | (Intercept) | -0.5410 | 0.1665 | -3.248 | <0.005 |
| | Validity - true | -1.6345 | 0.2983 | -5.480 | <0.001 |
| | Congr - incongruent | -0.2048 | 0.2293 | -0.893 | 0.371 |
| | Congr - neutral | -0.0803 | 0.2237 | -0.359 | 0.719 |
| | true:incongruent | 1.4373 | 0.3804 | 3.778 | <0.001 |
| | true:neutral | 0.4847 | 0.3980 | 1.218 | 0.223 |

$Model : IA \sim SentenceValidity * GazeCongruency + (1|subject) + (1|item),$
$family = binomial(link = "logit")$

suggests that people were more likely to inspect the linguistically identified object. That is, the target was inspected more frequently when the statement was true and the competitor was inspected less frequently when the statement was true compared to when it was false. The interaction of both predictors Sentence Validity and Gaze Congruency ($\chi^2(2) = 20.289; p < 0.001$) suggests that in congruent conditions people continuously inspected the object fixated and mentioned by the robot whereas in incongruent conditions participants made a visual attention shift from the object fixated by the robot to the object actually mentioned by the robot (cf. Figure 3.13. Figure 3.13 also shows that the condition true-neutral elicits similar responses as true-congruent, while false-neutral is similar to false-incongruent. The reason is probably that in neutral conditions participants have to rely on the linguistic information only and according to this, the target is the most likely referent until the sentence-final adjective is mentioned. This is also suggested by the inspection pattern in IP2, when people inspect the target rather than the competitor when gaze is neutral. Thus, in case of true statements, the utterance confirms the hypothesis that the target is the actual referent (hence the similarity to true-congruent) whereas mentioning the competitor (false statement) is inconsistent with previous hypotheses (hence the similarity to false-incongruent).

### Response Times

Trials were excluded from response time analysis when participants gave a wrong answer (4%) or response time was $\pm 2.5 \times SE$ above or below a subject's mean (1.69 %). Model reduction on the remaining data suggests that both predictors, Sentence Validity and Gaze Congruency, contribute to fitting a model to the data (Validity: $\chi^2(1) = 19.06; p < 0.001$, Congruency: $\chi^2(2) = 60.43; p < 0.001$). Model simplification further suggests that the interaction of the two predictors is marginally significant ( $\chi^2(2) = 5.598, p = 0.061$) but with more degrees of freedom and a higher BIC (15,897.6 vs 15,889.3 of the model without interaction) it is questionable which is the best model. We include a summary of the model containing the interaction in Table 3.6 along with p-Values obtained by MCMC-sampling (a negative coefficient reveals a shorter response time of the given level compared to the intercept level). Participants were significantly faster in responding when they had to give a positive answer (true condition) than when the robot's utterance was false. Moreover, people were also significantly faster in congruent trials, that is when the robot's gaze and utterance referred to the same object, compared to when robot gaze was neutral or incongruent. Reorganizing predictor levels within the model reveals that neutral and incongruent gaze do

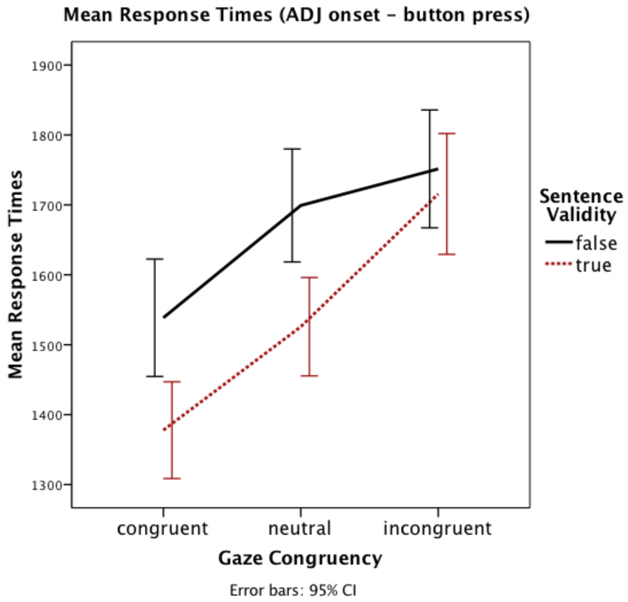


Figure 3.14.: Average response times for true and false statements, per gaze congruency condition.

not differ significantly in the elicited response times, but the numerical tendency for increased response time in incongruent trials is clearly visible. The reason for the lack of significance may be related to the difference between true-neutral and false-neutral conditions. As already suggested for the analysis of IP3-inspections, true-neutral and true-congruent similarly elicit and confirm the correct hypothesis, while false-neutral and false-incongruent conditions both initially elicit inspections to the target and then confront participants with conflicting information by identifying the competitor. It is, thus, not surprising that the difference in response times between false-neutral and false-incongruent is relatively small while the difference between true-neutral and true-incongruent is relatively large. In fact, it is noteworthy that true-congruent is significantly faster than true-neutral (according to post-hoc pairwise comparison with $p < 0.01$) since linguistic constraints select the target in both cases. It seems that robot gaze is such a strong, assuring cue that participants have an even stronger hypothesis about the validity of the sentence and, thus, can respond faster.

The effect of Gaze Congruency on response times suggests that people continuously follow and use robot gaze for utterance comprehension. Specifically, the facilitation ef-

Table 3.6.: Summary of model fitted to RT data.

| Predictor | Coefficient | *SE* | t-value | |
|---|---|---|---|---|
| (Intercept) | 1546.288 | 55.607 | 27.808 | |
| Validity - true | -166.874 | 48.744 | -3.423 | |
| Congr - incongr. | 203.771 | 48.699 | 4.184 | |
| Congr - neutral | 155.575 | 47.504 | 3.275 | |
| true:incongruent | 139.539 | 69.262 | 2.015 | |
| true:neutral | -4.033 | 68.152 | -0.059 | |

| | Coefficient | MCMCmean | pMCMC | Pr(> $\|t\|$) |
|---|---|---|---|---|
| (Intercept) | 1546.288 | 1546.103 | 0.0001 | <0.001 |
| Validity - true | -166.874 | -167.243 | 0.0002 | <0.001 |
| Congr - incongr. | 203.771 | 204.294 | 0.0002 | <0.001 |
| Congr - neutral | 155.575 | 155.603 | 0.0008 | <0.005 |
| true:incongruent | 139.538 | 138.640 | 0.0470 | <0.05 |
| true:neutral | - 4.033 | -4.043 | 0.9574 | 0.953 |

$Model : RT \sim SentenceValidity * GazeCongruency + (1|subject) + (1|item)$

fect of congruent gaze suggests that people used the gaze direction to anticipate potential referents such that incongruent gaze led to wrong expectations, resulting in slower response time.

One might argue that people followed and used robot gaze only until they realized that robot gaze was not so beneficial overall. Thus, the reported effects could have been very large in the first part of the experiment, but then disappeared in the second part when people had learned to ignore gaze because it was not generally helpful. This argument presupposes that people *decide* whether or not they interpret robot gaze movements as shifts in the robot's visual attention and whether they follow them or not, depending on the utility for task completion. It is indeed vaguely possible that the overall effects reported above are carried by behavior occurring in the first part of the experiment only. To investigate whether participants learned to ignore gaze or whether the reported effects are persistent for the entire experiment, we conducted a block analysis. An additional (binary) predictor "Block" captures the fact whether an item, and the produced response time, occurred in the first or in the second experimental block. Since items and conditions were sequentially randomized, each condition appeared twice on average in each block. Even though results from analyses with such sparse data are only indicative, they describe a trend for either change or continuity in the observed behavior. Thus, if participants indeed stopped using gaze in the second half of the experiment, we would expect to observe an effect of congruency in the first block, but not in the second block. Consequently, Sentence Validity may still have affected response time but the Gaze Congruency effect is then expected to disappear in block 2, inducing an interaction of the predictor Block with the predictor Gaze Congruency. In contrast, if participants remain attentive to robot gaze, the additional predictor Block is unlikely to interact with Sentence Validity or Gaze Congruency.

Interestingly, model reduction reveals a main effect of Block ($p < 0.001$, see also the sequential analysis of variance table in 3.7 such that we include it in the model as an additional predictor. However, we did not observe an interaction with other predictors. The final model is described in Table 3.8 and MCMC-sampling again confirms the main effect of Block. The negative coefficient for the second block indicates that participants are generally faster in the second half of the experiment. This is a frequent effect reflecting that people improve in task completion through practice. More importantly, the absence of an interaction suggests that participants similarly follow and use robot gaze throughout the entire experiment. Additional analyses within each block confirm this result and reveal that in block 1 as well as in block 2, Congruency

Table 3.7.: Model Reduction with all Predictors (including Block).

| Predictor | DF | SumSq | Mean Sq | F |
|---|---:|---:|---:|---:|
| Validity | 1 | 3,774,166 | 3,774,166 | 18.9 |
| Congruency | 2 | 12,703,156 | 6,351,578 | 31.8 |
| Block | 1 | 4,224,974 | 4,224,974 | 21.2 |
| Congruency:Block | 2 | 1,148,970 | 574,485 | 2.9 |
| Validity:Block | 1 | 129,444 | 129,444 | 0.6 |

$Model : RT \sim SentenceValidity * GazeCongruency * Block$
$+(1|subject) + (1|item)$

Table 3.8.: Final RT model including Block as a factor.

| Predictor | Coefficient | *SE* | t-value |
|---|---:|---:|---:|
| (Intercept) | 1605.87 | 56.96 | 28.194 |
| Validity - true | -166.09 | 48.28 | -3.440 |
| Congr - incongr. | 212.92 | 48.28 | 4.410 |
| Congr - neutral | 159.84 | 47.06 | 3.396 |
| Block - second | -125.49 | 28.04 | -4.475 |
| true:incongruent | 127.93 | 68.65 | 1.863 |
| true:neutral | 1.58 | 67.52 | 0.023 |

$Model : RT \sim SentenceValidity * GazeCongruency + Block$
$+(1|subject) + (1|item)$

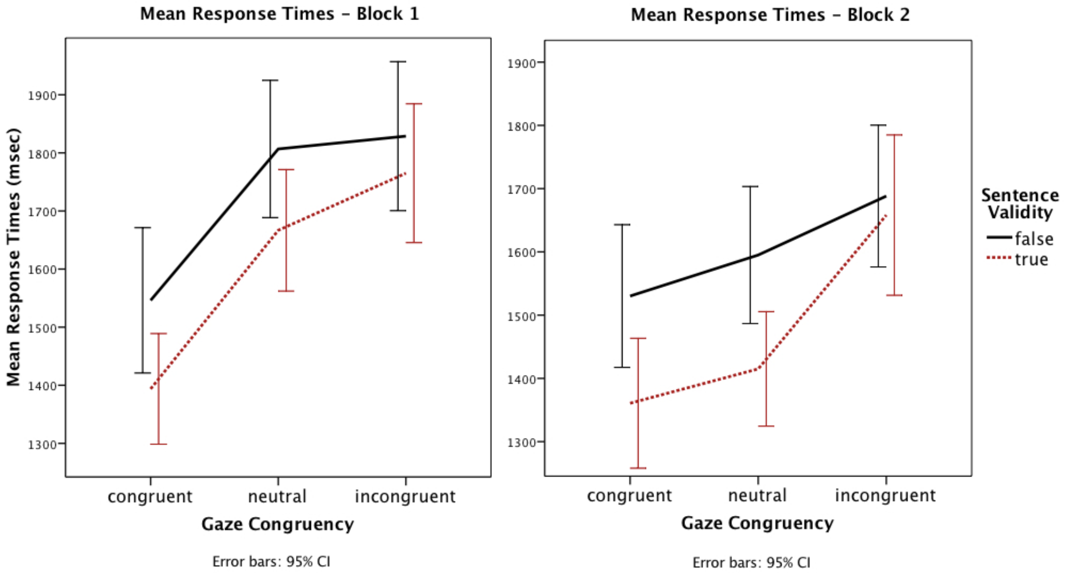


Figure 3.15.: Response times in block 1 (left graph) and block 2 (right graph).

and Validity have main effects. Notably, the response time for robot utterances with neutral gaze changes slightly across blocks. Fitting models separately to each block reveals that, in block 1, the facilitation effect of (congruent) robot gaze is most dominant, with levels *neutral* (Intercept) and *incongruent* not differing significantly from each other ($Coeff. = 75.80; SE = 51.83; t-value = 1.463; p = 0.14$). In block 2, the disruptive effect of incongruent gaze seems to have increased, i.e., the neutral condition elicits response time that is significantly slower than in the congruent condition ($Coeff. = -96.73; SE = 46.05; t-value = -2.101; p < 0.05$) but significantly faster compared to the incongruent condition ($Coeff. = 153.85; SE = 45.16; t-value = 3.407; p < 0.001$).

### Combined Analyses

We have analyzed two dependent variables, response time and inspection data, separately so far mainly because they have different properties. However, it appears reasonable to investigate the relation of these two dependent variables since they are both observed in response to the same manipulation of the stimuli. This way, one set of data can possibly help to examine reasons for the variation of another set of data. In our case, eye-movement data is observed as an on-line measure *during* exposure to the stimuli while response time is a measure recorded *after* perceiving the stimulus. Since our main manipulation concerned the robot's gaze direction (congruency) which occurred in the middle of a trial (as opposed to sentence validity which is a manipulation of the final part of a trial), participants' eye movements may potentially help to understand and explain how people's visual attention during a trial relates to their response time.

Recall that we found that the predictor Gaze Congruency affected response time, but in precisely what way remained speculation. We further found that participants followed robot gaze to an object and hypothesized that this visual referent may be considered to predict the linguistic referent. To shed some light on the relation between gaze-following and the response time effect, we included people's inspection behavior during IP1 (robot gaze towards target/competitor) as a predictor for a model of response time data. We predicted that, if the early visual cue to a potential referent led people to form a hypothesis about upcoming linguistic references, people who actually follow gaze would be faster in congruent trials. Similarly, following robot gaze to an object that was eventually not mentioned would mislead people and, thus, slow them down. In contrast, ignoring robot gaze and not looking at the visual referent is predicted to flatten this effect and result in a response pattern similar to the neutral gaze condition.

The data were coded as *following robot gaze* ('1') when participants had inspected the IA that the robot looked at at least once during IP1 and as *not following robot gaze* ('0') otherwise. Since we were interested in the effect of gaze-following, the neutral gaze condition was dropped in this analysis. The resulting data set includes subject and item information, the experimental condition (true/false, congruent/incongruent) as well as whether participants followed the robot gaze to the visual referent or not, and their response time. Model reduction shows that the predictor *GazeFollowed* interacts with *Gaze Congruency* ($\chi^2(3) = 11.425; p < 0.01$). The interaction introduces a larger BIC to the model but log-likelihood is largest, too, and since we are interested particularly in this interaction we include it in the final model summarized in Table 3.9.

Figure 3.16 depicts mean response times as a function of (a) whether people followed robot gaze (represented by lines "follow" versus "NOTfollow"), (b) whether robot gaze was congruent or not, and (c) whether the sentence was valid or not. Crucially, the interaction between GazeFollowed and Congruency which is also visible in Figure 3.16 suggests that facilitation as well disruption of robot gaze cueing a visual referent are larger when participants actually follow that cue and look at the potential referent. Participants that did not look at the visual referent showed smaller differences in their

Table 3.9.: Model summary and according p-Values from Markov chain Monte Carlo (MCMC) sampling.

| Predictor | Coefficient | *SE* | t-value |
|---|---|---|---|
| (Intercept) | 1572.89 | 62.49 | 25.171 |
| Validity - true | -156.69 | 62.62 | -2.502 |
| Congr - incongruent | 140.24 | 63.08 | 2.223 |
| GazeF - followed | -84.87 | 77.78 | -1.091 |
| true:incongruent | 102.91 | 90.32 | 1.139 |
| true:followed | -22.76 | 110.04 | -0.207 |
| incongruent:followed | 205.48 | 108.47 | 1.894 |
| true:incongruent:followed | 107.88 | 155.18 | 0.695 |

| | Coefficient | MCMCmean | pMCMC | Pr(> $\|t\|$) |
|---|---|---|---|---|
| (Intercept) | 1572.89 | 1570.95 | 0.0001 | <0.001 |
| Validity - true | -156.69 | -152.06 | 0.0206 | 0.013 |
| Congr - incongruent | 140.24 | 144.07 | 0.0212 | 0.027 |
| GazeF - followed | -84.87 | -83.70 | 0.2882 | 0.276 |
| true:incongruent | 102.91 | 95.55 | 0.2966 | 0.255 |
| true:followed | -22.76 | -31.23 | 0.7802 | 0.836 |
| incongruent:followed | 205.48 | 201.00 | 0.0742 | 0.059 |
| true:incongruent:followed | 107.88 | 116.37 | 0.4652 | 0.487 |

$$Model : RT \sim SentenceValidity * GazeCongruency * GazeFollowed$$
$$+(1|subject) + (1|item)$$

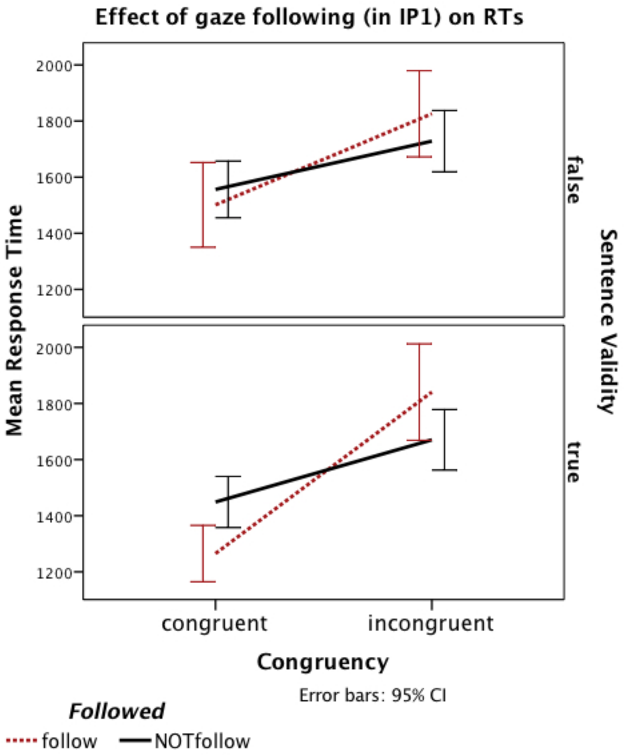


Figure 3.16.: Inspection pattern predicting response times. When people had followed robot gaze to the target/competitor in IP1, gaze congruency had a greater effect on response times.

response times. Interestingly, the main effect of congruency – even though smaller – remains which suggests that people did take notice of the visual referent, possibly covertly. This result does not establish a causal link but further supports the claim that robot gaze cues a visual referent which influences people's hypotheses about the utterance.

Concluding our results from Experiment 2, we find that the response time results are in agreement with our eye-movement data and suggest that participants follow both robot gaze and robot speech. The response time, and in particular its direct relation to robot-gaze-following, supports the interpretation that congruent gaze benefits and incongruent gaze disrupts comprehension. We argue that people follow robot gaze to an object and form hypotheses about what the robot is going to mention next, that is, robot gaze is interpreted with respect to referential intentions. When people's expectations about an upcoming referent are met, as is the case in congruent trials, people are faster to respond. However, when the visual and spoken references mismatch, the comprehension process seems to be disrupted and increased response time are elicited.

The participant behavior we observed in response to robot behavior is in many respects similar to what Hanna and Brennan (2007) observed in their studies. We similarly found that: (a) Listeners begin to orient visual attention in the same direction as the robot/speaker within 1,000ms after "VPoD" (visual point of disambiguation, i.e., the first speaker's look towards the referent before/while beginning to speak which corresponds to our robot's gaze onset). (b) Listeners follow the robot/speaker's gaze during scene and utterance comprehension. (c) Listeners use this gaze cue to early disambiguate an utterance with respect to the scene, that is, they look at the target rather than the competitor well before the LPoD. Note, that in Experiment 1 listeners similarly kept looked at the visual referent during IP2 suggesting that gaze eliminated referential ambiguity. We therefore conclude that people use robot gaze in a similar way that they use human gaze. That is, the observed fixation patterns in response to robot gaze are also consistent with and extend the idea that gaze elicits reflexive visuospatial orienting (Friesen and Kingstone, 1998; Driver et al., 1999; Langton and Bruce, 1999). The persistence of the observed congruency effects in particular seem to suggest that people *automatically* follow robot gaze. The reader is referred to Chapter 6 of this thesis for further discussion of reflexive attention shifts in response to gaze cues. However, the observed effects of robot gaze congruency on utterance comprehension in term of response time further suggest that people *use* the visual information provided by the robot's gaze.

Furthermore, we would like to point out that Hanna and Brennan reported in their studies that listeners rarely looked at the speakers' face to detect where the speaker was gazing at and rather used the speaker's head orientation peripherally. This is additional support for the claim that the type of robot gaze used in our studies – that is, as a combination of head and gaze movement – can in principle be used in much the same way that human speakers' gaze is used even though the robot has no anthropomorphic appearance and no human-like eyes. We suggest that it is sufficient for people to ascribe the function of 'seeing' to the camera in order to elicit similar behavior that human gaze elicits.

Finally, visuo-spatial orientation induced by (robot) gaze seems to constrain the domain for utterance interpretation which in turn affects reference resolution. That is, people seem to prepare to ground an upcoming referring expression with respect to one or several objects that have become the focus of visual attention (as a result of robot gaze). While this in line with Hanna and Brennan's results, their interpretation relied exclusively on eye-movement data. Our response time data now provide additional support for the effect of gaze cues on reference resolution. Thus, the presented evidence supports the hypothesis that people interpret robot gaze as a visual reference cueing the linguistic reference. This, however, suggests that people relate visual and linguistic cues to each other, possibly via the assignment of attentional states to robot gaze. The results presented in this Chapter therefore support the hypothesis, that people establish *joint attention* with the robot and that this can be beneficial for comprehension (response level 3, as introduced in Section 1.4).

However, the behavioral data from the presented experiments do not reveal in what way the knowledge about the visual referent (that the participants jointly attend to) affects utterance interpretation. Whether people not only attend to the same object but also draw inferences about *why* the robot attended to this object remains speculation. Two different mechanisms may explain the observed facilitation/disruption effect on response times: People may use robot gaze in a "top-down" manner, driven by the belief that robot gaze reflects attentional and intentional states, revealing what the robot intends to mention (we call this the *Intentional Account*). Thus, people possibly infer referential intentions from the robot's gaze so that the expectation of a referent facilitates (or, if incorrect, disrupts) comprehension. Alternatively, people follow robot gaze (reflexively) to the visual referent without inferring any communicative intentions. Robot gaze rather happens to direct their visual attention to the right (or wrong) referent at the right (or wrong) time such that further processing is influenced accordingly (*Visual Account*). To present a first attempt at distinguishing these two account, this issue was specifically addressed by Experiment 3.

## Summary

In Experiments 1 and 2, we observed that people follow both the robot's gaze and utterance to referents in the scene. In Experiment 1, the robot ambiguously referred to two potential referents and we observed that people inspected the object looked at by the robot significantly more frequently than the other potential referent. In fact, people followed robot gaze both when the sentence uniquely or ambiguously identified a scene referent. Response times were similar for the referentially ambiguous utterance compared to an unambiguous utterance, suggesting that gaze eliminated referential ambiguity. In Experiment 2, we used globally unambiguous utterances and investigated whether robot gaze is indeed beneficial for utterance comprehension such that people validate an utterance faster than when gaze is neutral. To separate the influence of robot gaze and speech, we manipulated the *congruency* of robot gaze and speech such that gaze sometimes served as a cue to a subsequently mentioned object and sometimes it did not. Moreover, we manipulated the *validity* of the statements which were unambiguous in general but contained temporarily ambiguous referring expressions. We again observed gaze-following in all conditions, that is, people inspected the visually cued object more frequently than the potential competitor. People inspected this object until the same object or the competitor object was uniquely identified by the robot's mention of the color of the respective object. While congruent robot gaze was observed to speed participants' response time, a mismatch between visually cued and linguistically identified object led to slowed response time. These results suggest that people interpreted the gaze cue to indicate which object was going to be mentioned by the robot. When people's expectations with respect to upcoming referents were met, response times were faster compared to when people had to revise these expectations and re-validate the utterance.

These results support the assumption that people apply response level 3 (introduced in Section 1.4) during interaction with our robot. That is, participants establish *joint attention* with the robot and the acquired visual information facilitates comprehension. However, the presented results do not reveal how the object in visual focus influences comprehension. Conceivable are two different mechanisms: (a) People infer referential intentions from robot gaze which facilitates comprehension when these intentions are congruent with the actual utterance, or else this disrupts comprehension (*Intentional Account*), or (b) people (possibly reflexively) attend to the visual referent but do not infer any communicative intentions and the late attention shift in incongruent trials simply requires additional effort and time (*Visual Account*). Thus, whether people indeed form beliefs on the robot's communicative intentions, or whether the increased effort related to an additional visual attention shift simply increases response time is further explored in Chapter 4.

# 4. Inferring Referential Intentions from Gaze

The results of Experiments 1 and 2 suggest that robot gaze is a strong cue which guides visual attention in an automatic fashion and that this further influences utterance comprehension. However, there are two categories of explanation for the observed response time effects. Either people infer referential intentions from the robot's gaze so that the expectation of a referent facilitates (or, if incorrect, disrupts) comprehension (called the *Intentional Account*). Or, people jointly attend to the visual referent but do not infer any communicative intentions and robot gaze simply induces a visual attention shift either to the right object at the right time (facilitation) – or not (disruption). We call this the *Visual Account*.

If, indeed, the facilitation/disruption effect of robot gaze on people's comprehension is due to the inferred referential intentions, we predicted that robot gaze would not only affect how fast references are resolved but also which object is believed to be the intended referent of the utterance. Such behavior would provide strong evidence supporting the Intentional Account, or response level 3 (Section 1.4), and some indication for response level 4. That is, if robot gaze was shown to affect beliefs about which referent the robot *intended* to talk about, this would clearly suggest that people interpret robot gaze with respect to attentional and intentional states (leading to *joint attention*, level 3) and the inference of referential intentions could even indicate that response level 4 (*shared attention*) is possible in HRI.

## 4.1. Experiment 3

Experiment 3 more thoroughly investigated how robot gaze affects reference resolution when people have to correct the robot utterance. A verbal correction implicitly requires participants to actively decide which referent they think was intended by the robot, avoiding the need to explicitly ask people for this decision. The user, thus, is engaged in a task designed to reveal the relative importance of linguistic and gaze cues for identifying an intended referent.

Recall the example scenario described for Experiment 2: Person A wants an ingredient but cannot reach for it since her hands are covered in cup cake dough. So she asks person B for help. There is a bowl of macadamia nuts to A's left and a bowl of hazelnuts to her right. But A is in a rush and confuses things, so she says: "B, please pass me the hazelnuts" while looking at the macadamias. In this case it is not obvious that the sentence is false but the referential cues are incongruent. Thus, B may decide which type of nuts A meant to ask for and try to pass those, or she can correct and clarify: "A, you wanted the hazelnuts, right?" if she thinks that A just looked at the wrong bowl, or "A, you wanted the macadamias?" if she thinks that A just used the wrong name.

In our experiment, we observed participants mainly in response to false robot utterances which required a verbal correction. These were, as in Experiment 2, accompanied by either congruent, incongruent or neutral robot gaze. Thus, when the robot mentioned one object but gazed at another, participants needed to produce a correction that involved the object they considered to be the intended referent. The aim of this experiment was to determine whether and to which extent robot gaze modulates participants beliefs about referential intentions. Additionally this experiment served to see whether the previously observed visual attention pattern in participants was robust to changes in the task and could be replicated. A post-experimental questionnaire further sought to assess the general beliefs and impressions people obtained from the interaction with the robot.

### 4.1.1. Method

#### Participants

Thirty-six native speakers of German, again mainly students enrolled at Saarland University, took part in this study (12 males, 24 females). All reported normal or corrected-to-normal vision.

#### Materials

For Experiment 3, we used the same set of stimuli that was used for Experiment 2. That is, 24 items were used which occurred in six conditions each. The conditions resulted from the manipulation of Sentence Validity (true/false) and Gaze Congruency (congruent, incongruent, neutral; see Table 3.2). Because we wanted to mainly analyze the correction statements participants produced, false robot utterances were of particular interest in this experiment. Using the previous example sentence and scene, such a

Table 4.1.: 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* | *Color* |
| false - neutral: | — | Target | Competitor |
| false - congruent: | Competitor | Target | Competitor |
| false - incongruent: | Target | Target | Competitor |

false utterance "The cylinder is taller than the pyramid that is brown" would be accompanied either by robot gaze towards the brown pyramid (congruent), the pink pyramid (incongruent) or neutral gaze. In those utterances, there were two cues identifying the referent. The first cue was the comparative (*taller than* or *shorter than*) and the second cue was the object color. False statements were false when these two cues did not identify the same referent, e.g., when the cylinder was not *taller* than the *brown* pyramid. Thus, people could repair such an utterance by changing either the comparative or the color adjective in their correction sentence.

The neutral condition provided a baseline concerning the bias towards either repair option in the absence of gaze, that is, wether people generally preferred to adapt the comparative, for instance. When referential robot gaze was present it emphasized one of the potential referents: either it supported the mentioned object (identified by color) or it supported the alternative object (identified by the comparative, not color). Details on referential variation for the three *false* conditions are shown in Table 4.1.

### Procedure

In this experiment, people were instructed to give an oral correction of the robot's utterance when they thought that the robot had made a mistake. This formulation was deliberately kept rather vague so that people were free to interpret "mistake" in a way they found appropriate. The "cover story" for this experiment remained the same as in Experiments 1 and 2, i.e., participants were told that the robot system would be evaluated. It still made many mistakes and participants' feedback was to be used as feedback in a machine learning procedure to improve the robot system. They were further told to start their correction with the same object reference that the robot started with, making it easier for the system to learn from the corrected sentences. Once more, this explanation served as a cover story making the task appear plausible. People's utterances were recorded from trial start to end, that is from video onset until participants pressed a button indicating that they finished giving their correction. Thus, the experiment was self-paced and participants could start their utterance at any time during a trial. Participants' sentences were recorded using a mobile microphone connected to an Asio AudioCard. The eye-tracker adjustment and calibration procedure as well as drift correction and presentation of the stimuli were otherwise identical to Experiments 1 and 2.

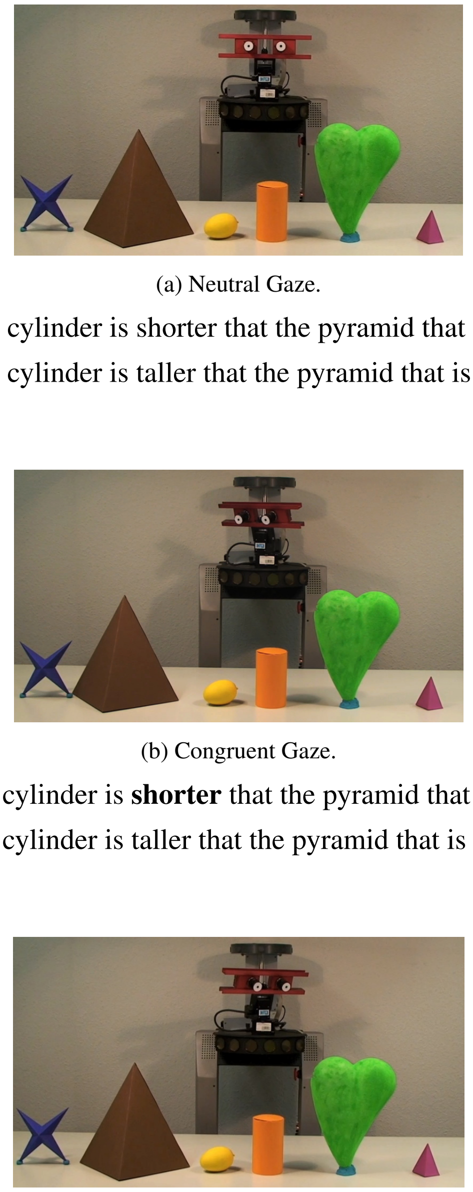
### Analysis

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 4.1). The three categories assigned to responses were *Target* (object matching the comparative), *Competitor* (object matching the color adjective) and *Else* (no correction given or described one or more different objects). Each response category was thus coded as a binary variable (e.g. the target had been described in the correction sentence ('1') or not ('0')). Since participants almost always either produced a sentence containing the target or the competitor, both response categories *Target* and *Competitor* were nearly complementary. While we consider only false utterances and removed Sentence Validity as a factor, the remaining predictor 'Gaze Congruency' has again three levels: congruent, neutral and incongruent. For the analysis, we used logistic regression similar to the mixed-effects models used for eye-movement data.

We again recorded people's eye movements during trials in order to compare participant behavior in this study with the behavior observed in previous studies. The analysis of eye-movement data was identical to Experiment 2.

### Predictions

In false utterances, the target object (which is consistent with the comparative of the sentence) is not the one identified at the LPoD (color adjective). That is, in the false-congruent condition, the robot looks at the competitor which is also identified at LPoD but is not consistent with the comparative. In the false-incongruent condition, the robot



(a) Neutral Gaze.

i) The cylinder is shorter that the pyramid that is brown.

ii) The cylinder is taller that the pyramid that is pink.

(b) Congruent Gaze.

**i)** The cylinder is **shorter** that the pyramid that is brown.

ii) The cylinder is taller that the pyramid that is pink.

(c) Incongruent Gaze.

i) The cylinder is shorter that the pyramid that is brown.

**ii)** The cylinder is taller that the pyramid that is **pink.**

Figure 4.1.: Predicted corrections of the sentence *"The cylinder is taller than the pyramid that is brown."* for neutral, congruent and incongruent robot gaze. In i) the comparative is adapted to match the competitor object, in ii) the color adjective is adapted to match with the target object.

looks at the target, which is not identified at LPoD but is consistent with the comparative. Thus, people may correct the robot's utterance while mentioning the competitor (and changing the comparative) or they could produce a correction containing the target (by changing the LPoD, color adjective). The neutral condition shows whether there is a bias towards changing one of the two linguistic cues (comparative versus adjective) and provides a baseline to compare against robot-gaze induced repair preferences. In Figure 4.1, the possible correction sentences are given as well as a predicted preference in each gaze condition.

We previously hypothesized that the effect of robot gaze on people's comprehension was due to the assumption that robot gaze reflects attentional states such that it, consequently (and similar to human gaze), elicits predictions about the intended referent of the speaker. If this Intentional Account holds, we predicted that robot gaze would not only affect how fast references are resolved (Experiment 2) but also which object is understood to be the referent. More precisely, we would expect people to describe the target and adapt the color adjective, for instance, more often in the false-incongruent condition (when the robot looks at the target, Figure 4.1 c) than in the false-congruent or false-neutral conditions – even though the (false) utterance always identifies the competitor at LPoD. Similarly, we predicted that people would tend to describe the competitor and change the comparative accordingly when the robot also looked at the competitor (false-congruent, Figure 4.1 b). If robot gaze, however, directs the listener's visual attention towards an object without contributing referential meaning (Visual Account), a significant difference in people's repair patterns across the three gaze-conditions would be unlikely.

### 4.1.2. Results and Discussion

#### Eye movements

Plots in Figure 4.2 and 4.3 show a fixation pattern that is extremely similar to that observed in Experiment 2 (Figures 3.9, 3.10). This suggests that findings on people's visual attention during this experiment replicate the findings from our previous experiments. That is, people robustly followed the robot's gaze and speech to objects in the scene irrespective of the type of task they were given.

Inferential analyses of the respective IPs confirmed that people reliably followed robot gaze. In IP1, people inspected the target more frequently when the robot looked at the target, i.e., in condition "targetgaze" combining true-congruent and false-incongruent, compared to when it looked at the competitor (Coeff.=$-2.78$, *SE* = 0.32, Wald Z =$-8.64$) or when gaze is neutral (Coeff.=$-2.27$, *SE* = 0.27, Wald Z = $-8.33$). Similarly, people were more likely to inspect the competitor when the robot looked at the competitor than when it looked at the target (Coeff.=$-3.27$, *SE* = 0.39, Wald Z = $-8.34$) or is neutral (Coeff.=$-2.21$, *SE* = 0.27, Wald Z =$-8.29$). The same pattern is observed for IP2, suggesting that people continued to inspect the object that had previously been looked at by the robot, even though the referring expression was ambiguous at that point. More precisely, in IP2, people were more likely to look at the target when the robot had fixated the target prior to noun onset, and the competitor was more often inspected when it was looked at previously by the robot.

Overall, the change from a response time task to a self-paced correction task did not seem to affect how robot gaze influences people's visual attention. Instead, people followed robot gaze in both settings - both under time pressure (Experiment 2) as well as in a self-paced correction task (Experiment 3). On one hand, the argument that people follow robot gaze as part of a strategy in order to better and faster fulfill the task is unlikely since robot gaze was neither generally helpful for task completion nor was there a need to respond particularly fast. On the other hand, it is highly unlikely that people follow gaze (and do not look at the robot head) purely for reasons of boredom or curiosity since gaze was frequently misleading and disrupting people in fulfilling the task of Experiment 2. Thus, the replication of the eye-movement results further support the view that people attend to (robot) gaze so closely and reliably, possibly because they reflexively react to robot gaze and further consider it to reflect attentional states.

### Sentence Production

Since participants had to start a correction sentence with the same object as was used in the original sentence (anchor), we mainly found corrections that additionally involved either the target or the competitor object. To assess whether the robot gaze cue influenced the choice of the object involved in a correction and whether an object itself elicited preferences for including it in a description, we initially included two predictors, Described Object and Gaze Congruency, in our analyses. Model reduction suggests that both predictors contribute to fitting a model to the data since their interaction was significant ($\chi^2(4) = 58.12; p < 0.001$). With more degrees of freedom, the log-likelihood of such a model is also largest (-541.06) and AIC and BIC are smallest (BIC of 1160.96 versus other models with a BIC of 1176.08, 1190.41 or 1685.68). A summary of the resulting model containing both predictors and the interaction is given in Table

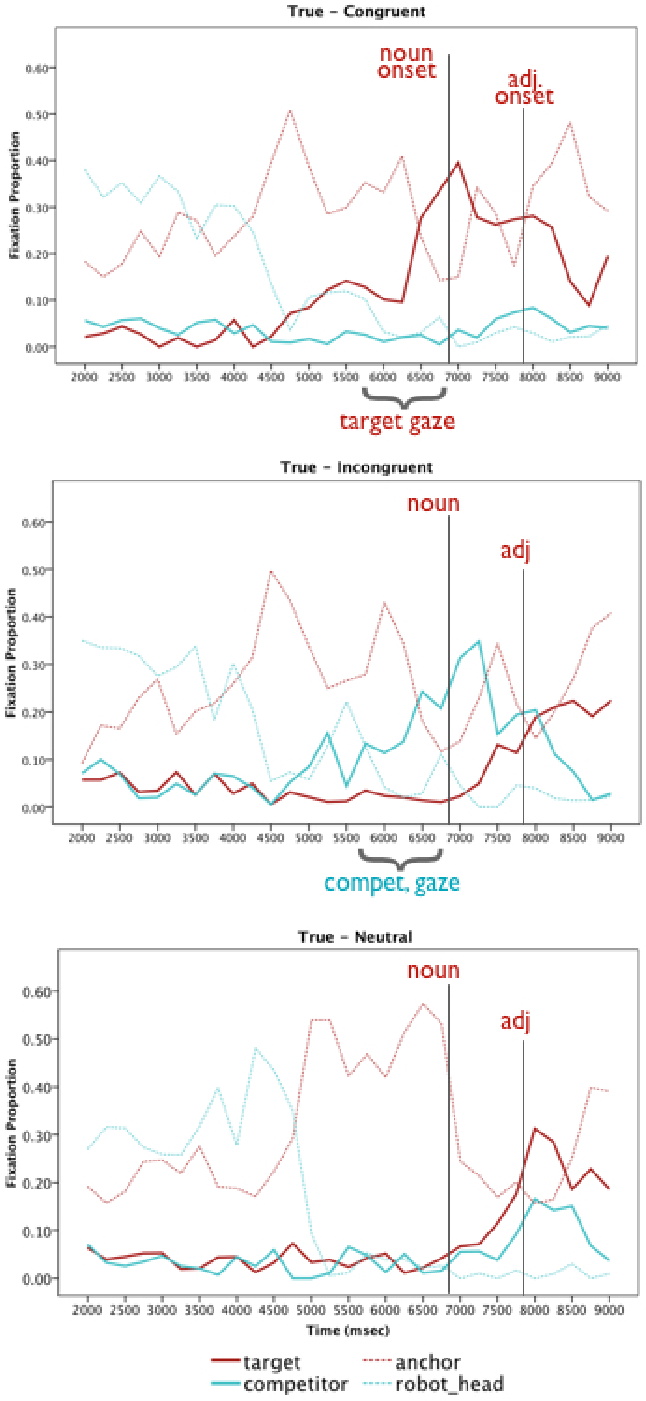


Figure 4.2.: Average Fixation proportions in 250ms bins across a whole trial.

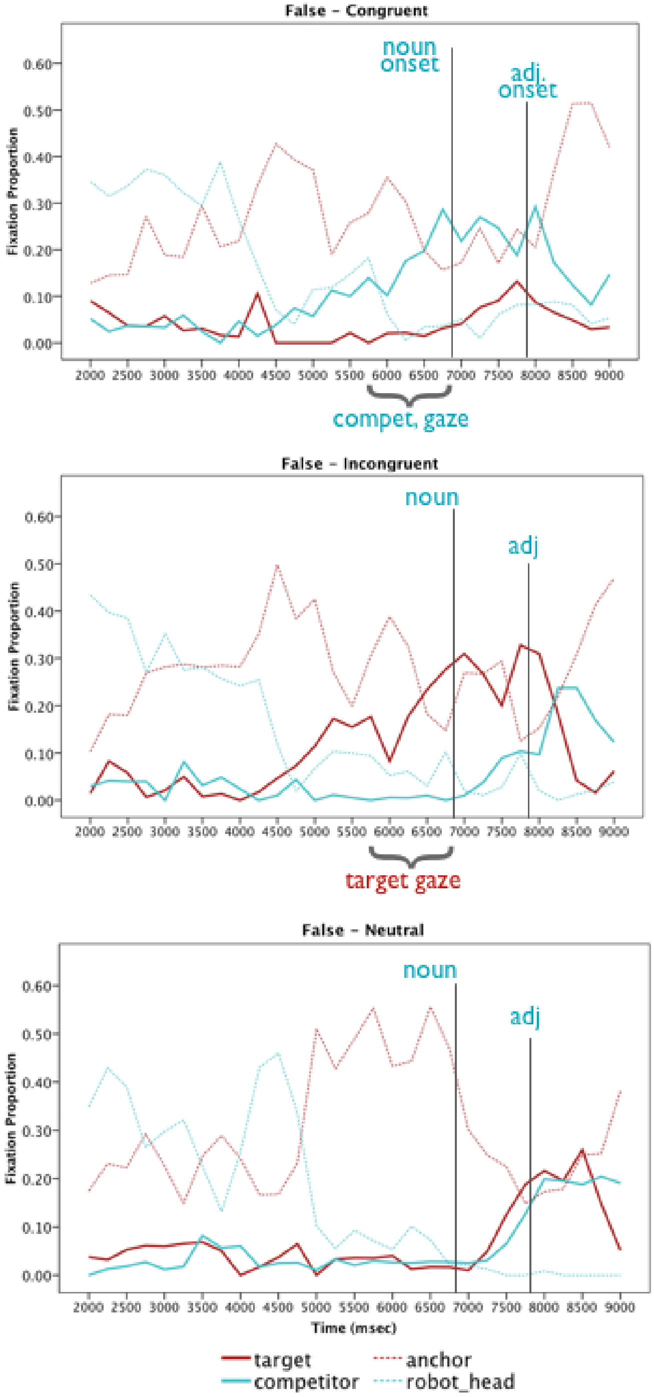


Figure 4.3.: Average Fixation proportions in 250ms bins across a whole trial.

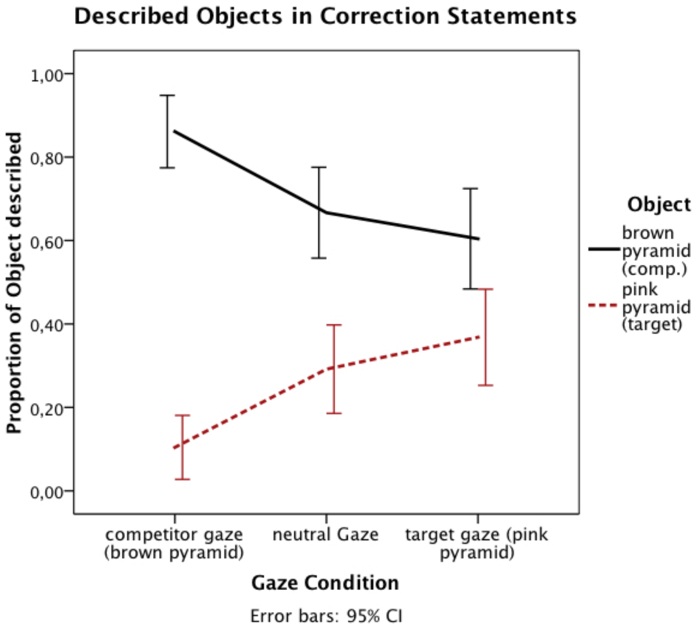


Figure 4.4.: Proportion of objects described in response to false robot utterances of the form *"The cylinder is taller than the pyramid that is brown"*.

4.2. Moreover, this table shows models for each response category individually indicating how well Gaze Congruency predicts when the target (or the competitor) is chosen. Since only false statements are considered in this analysis, the gaze condition *congruent* shows competitor gaze while the *incongruent* condition consequently shows target gaze.

The response category *Else* was found in 3.47% of the *false*-trials and was treated as missing values in the analysis. The reason for excluding category Else is that it is conceptually not a third response category. Instead, response category remains a binary (or dichotomous) dependent variable to which simple logistic regression can be applied. The alternative would be multinomial logistic regression for a polytomous dependent variable which is less interpretable and, at this stage, not available (Barr, 2008). We, thus, fitted a logistic regression model to our data including Response Category as predictor, specifying which object is part of a given description. This serves the purpose to find out whether one object was described generally more often than the other. Otherwise, we fitted simple logistic regression models for each response category

Table 4.2.: Summary of the resulting model (Model1) and summaries of models for separate outcome categories (success & failure of target/competitor).

| Predictor | Coefficient | *SE* | Wald Z | p |
|---|---:|---:|---:|---:|
| (Intercept) | 1.8245 | 0.241 | 7.572 | <0.001 |
| Object-target | -3.9763 | 0.364 | -10.924 | <0.001 |
| Congr-incongr. | -1.4017 | 0.295 | -4.749 | <0.001 |
| Congr-neutral | -1.1314 | 0.299 | -3.786 | <0.001 |
| target:incongr. | 3.0129 | 0.438 | 6.887 | <0.001 |
| target:neutral | 2.3959 | 0.444 | 5.393 | <0.001 |
| *success of described object 'target'*<br>(Model2) | | | | |
| (Intercept) | -3.1050 | 0.451 | -6.889 | <0.001 |
| Congr-incongr. | 2.2738 | 0.395 | 5.763 | <0.001 |
| Congr-neutral | 1.7608 | 0.395 | 4.454 | <0.001 |
| *success of described object 'competitor'*<br>(Model3) | | | | |
| (Intercept) | 2.6273 | 0.415 | 6.332 | <0.001 |
| Congr-incongr. | -2.0161 | 0.362 | -5.566 | <0.001 |
| Congr-neutral | -1.6179 | 0.363 | -4.461 | <0.001 |

$Model1 : UsedInAnswer \sim DescribedObject * GazeCongruency$
$+(1|subject) + (1|item), family = binomial(link = "logit")$
$Model2 : Target \sim GazeCongruency + (1|subject) + (1|item),$
$family = binomial(link = "logit")$
$Model3 : Competitor \sim GazeCongruency + (1|subject) + (1|item),$
$family = binomial(link = "logit")$

(target/competitor) separately accounting for the fact that the response categories are not independent. While the results from the inferential analyses are provided in Table 4.2, we also computed mean proportions of corrections involving the target/competitor and plotted them for visualization purposes in Figure 4.4.

In almost 67% of their correction statements in the neutral gaze condition participants preferably gave this correction sentence: "The cylinder is *shorter* than the pyramid that is brown." That is, in the neutral baseline condition we observed a general preference to build a corrected sentence involving the competitor (which has been linguistically identified by the mentioned color in false trials), changing the comparative accordingly. This is depicted in the central condition in Figure 4.4 and confirmed by the fixed effect of predictor Described Object in Model1, Table 4.2. The overall preference to keep the more explicitly mentioned object (color match) remained dominant in all three gaze conditions is likely due to two possible reasons. Firstly, gaze is frequently incongruent in our stimuli (and often considered incorrect) whereas speech is always fluent and clear. This may have induced a general bias to trust the competence of language rather than the gaze cue. Consequently, linguistic referential cues were preferred information for the identification of the intended referent (while gaze cues "only" modulated this process). Secondly, it has been shown that people prefer to use absolute (shape and color) to relative features (size, location) for the production of REs (Beun and Cremers, 1998). That is, among the linguistic referential cues, color was simply the most dominant cue to an intended referent.

A positive coefficient of the predictor Gaze Congruency in Model2 and Model3 is interpreted as a larger probability of describing the according object in a given predictor level. The results in Table 4.2 therefore indicate that people corrected an utterance using the target (i.e. change the color mentioned in robot statement) significantly less often when robot gaze was directed towards the competitor (false-congruent) compared to when the robot actually looked at the target (false-incongruent) or robot gaze was neutral (false-neutral). These results are depicted by the dotted line in Figure 4.4. Similarly, participants chose to give a scene description involving the competitor (and changing the comparative) with significantly higher probability when the robot's gaze was directed towards the competitor compared to when it was target-bound or neutral. This result is depicted by the continuous line in Figure 4.4. That is, the robot's gaze increases the likelihood of correcting the competitor or target in the congruent or incongruent condition, respectively.

Another observation suggesting that gaze did affect reference resolution becomes apparent when analyzing corrections in response to *true* robot utterances. Although we did not expect participants to correct true statements, interestingly, we observed that in 15% of true-incongruent trials (i.e., in 21 trials) people corrected the robot utterance with a sentence describing the competitor which the robot had looked at (true-congruent was corrected in 4 trials, true-neutral only in 1 trial). This suggests that people believed that the robot was indeed talking about the competitor, which it looked at, even though both the comparative and the mentioned color uniquely identified the target object. This result is surprising given that a task requiring sentence correction should induce a clear focus on the utterance. In the mentioned true-incongruent trials, however, participants most likely did not see the target otherwise they would have realized that the utterance was in fact correct. Instead, they must have focused completely on the object that the robot had fixated (competitor) leading to the unnecessary production of a correction sentence involving the competitor.

The results from the correction analysis support our hypothesis that people consider robot gaze to reflect its attentional state and, further, its intention to talk about an object that it looks at. These findings also suggest that people in fact establish joint attention and integrate the *visual reference* derived from robot gaze with their on-line interpretation of robot speech. Thus, these observations clearly favor the Intentional Account (response level 3, and possibly 4) for explaining the facilitation/disruption effects of robot gaze on comprehension reported in Chapter 3.

### Questionnaire

In order to examine what general impressions participants obtained from the robot and, in particular, whether the video-based presentation mode posed a problem for people, we asked participants to fill out a post-experiment questionnaire. This questionnaire contained questions and statements concerning people's general impression of the robot, whether they thought its utterances were comprehensible at all or what they thought were the robot's most common errors. Overall, 20 statements or questions had to be evaluated and responses were given on a 7-level Likert scale to allow for graded agreement (1=no/don't agree and 7=yes/agree). The results of this questionnaire are shown in Table 4.3. The first batch of statements indicates more generally what an impression participants obtained during the experiment from the robot. While the robot was not perceived as very natural or clever, people also did not find it especially confusing or annoying. Considering the persistent incongruent behavior and the frequent mistakes of the robot, this makes people appear rather patient. Question 6 and 8 are particularly important for the justification of our design, as we presupposed that the robot's utterances were easily assessable, despite the temporary ambiguity in the sentences or the video presentation allowing only a certain scene perspective. The mean agreement values suggest that participants found robot utterances comprehensible and that video presentation did not disrupt task completion. Interesting, and remarkably accurate, are also participants' judgements concerning most frequent errors of the robot: size comparisons as well as gaze behavior were considered the most frequent mistakes (number 14, 17). On the other hand, participants also thought that the robot typically looked at what it talked about. Participants further seemed to think that they hardly attended to the robot's movements (and, thus, did rather not follow the robot's gaze) as the mean of 3.3 in statement number 11 suggests. Notably, we observed clear and robust gaze-following in this and the previous experiments. That is, this result hints at the automatic or sub-conscious manner of the observed gaze-following behavior. Additionally, we were interested to see whether participants had hypotheses concerning the robot's competences. That is, statements number 19 and 20 sought to reveal whether people assumed speech production or visual processing to be the source of most errors. The means of 3.1 and 4.7 respectively are significantly different (t-test(speech,vision):$p < 0.001$) and indicate that participants rather trusted the robot's speech competence (e.g., it produced the correct color word or the correct comparative) than its ability to grasp the visual scene (it did not correctly recognize the size difference). On one hand, the frequently inappropriate camera movements – while speech was always fluent – may have influenced this result. On the other hand, people may generally think that correct scene comprehension is more difficult for an artificial system than speech production, possibly because lexical retrieval and sentence production appear easy and natural to themselves while recognizing and comparing various geometric shapes are not always easy for some participants.

### Summary

To distinguish the purely visual component of robot gaze from its potentially referential meaning, we changed the task from a response time task (Experiments 1 and 2) to a production task in Experiment 3. People had to verbally correct the robot's statement in a self-paced manner. False robot utterances contained two conflicting referential cues, the comparative and the LPoD with the color adjective, which selected either the target or the competitor, respectively. Additionally, robot gaze provided a referential cue

Table 4.3.: Statements and Average Agreement - from 1 (no) to 7 (yes))

| Nr. | Topic | Statement | Mean Agreement | SD |
|---|---|---|---|---|
| 1 | Impression | The robot's behavior is confusing. | 3.10 | 1.37 |
| 2 | | The robot's behavior is natural. | 3.23 | 1.48 |
| 3 | | The robot's behavior is clever. | 4.17 | 1.49 |
| 4 | | The robot's behavior is erroneous. | 4.47 | 1.31 |
| 5 | | The robot's behavior is annoying. | 2.07 | 1.55 |
| 6 | Clarity | Are the robot's utterances comprehensible? | 5.83 | 1.39 |
| 7 | Gaze&Head | Do you find the head movement appropriate and natural? | 3.53 | 1.61 |
| 8 | Presentation | Do you think task completion was more difficult due to video presentation? | 1.80 | 1.37 |
| 9 | Behavior | The robot is an intelligent system which rarely makes mistakes. | 4.30 | 1.64 |
| 10 | | The robot mostly looks at what it describes. | 4.57 | 1.76 |
| 11 | | I did not pay attention to the robot and and concentrated on the sentences. | 3.30 | 1.51 |
| 12 | | I can imagine a natural conversation with it. | 2.37 | 1.25 |
| 13 | | I don't think the robot is intelligent, and its utterances are rather random. | 2.57 | 1.19 |
| 14 | Errors | The robot mostly made false size comparisons. | 4.37 | 1.50 |
| 15 | | The robot mostly named the wrong shape. | 2.77 | 1.28 |
| 16 | | The robot mostly did not recognize color correctly. | 4.40 | 1.38 |
| 17 | | The robot mostly looked at a wrong object. | 4.60 | 1.52 |
| 18 | | The robot mostly named the wrong locations. | 3.33 | 1.32 |
| 19 | Error Source | Mostly Speech Production: Robot often chooses incorrect words. | 3.13 | 1.74 |
| 20 | | Mostly Vision Processing: Robot often does not correctly recognize scene entities. | 4.67 | 1.52 |

to either target or competitor. Which referent (target or competitor) participants mentioned in their correction sentences was, thus, assumed to reflect which referent they thought was intended by the robot. Interestingly, the produced correction statements revealed that robot gaze modulated which referent people mentioned in their correction sentence, i.e., which object they understood to be the intended referent. Since this experiment imposed no time pressure on people's responses, the effort of shifting visual attention alone cannot account for effects of robot gaze on comprehension and response production. People had sufficient time to reorient their visual attention and prepare their utterance. Rather, the referential intention inferred from robot gaze seems to have influenced what objects people selected as referents for their corrections. In the light of these results, the Visual Account can probably be rejected as explanation for facilitation/disruption effects of robot gaze (Experiment 2) and preference modulation of referent selection for correction statements (Experiment 3).

Thus, in Experiments 1-3, we have provided evidence that listeners not only follow robot gaze but that they further use this as a visual reference revealing communicative intentions. In these experiments, we assumed that robot gaze would be only beneficial for utterance comprehension if it was aligned with speech in a human-like manner such that gaze cues could be similarly used as referential cues. We argued that human-like speaker gaze facilitates communication because it fulfills certain functions in communication but, in fact, it is an open question whether such alignment is indeed required for people to interpret robot gaze as an attentional and intentional cue that affects how people comprehend the utterance. The investigation of the relevance of alignment between gaze and speech for maintaining its utility during utterance comprehension further examines whether the Visual or the Intentional Account is most likely to explain the influence of robot gaze. In the following chapter, we thus consider these issues in greater detail.

# 5. Synchronization of Gaze and Speech

Experiments 1-3 have examined the influence of robot gaze, when it is aligned with speech in a human-like manner. Specifically, in Experiments 1 and 2, we have shown that congruent gaze which is synchronized with speech not only elicits gaze-following but also facilitates comprehension. In contrast, incongruent gaze has been shown to disrupt comprehension such that listeners needed more time to validate a sentence than when gaze was neutral. We hypothesized that the facilitating/disruptive effect of congruent/incongruent robot gaze (respectively) was either due to a helpful or a useless reflexive shift of visual attention (Visual Account), or it was due to people interpreting robot gaze as a cue to its attentional and intentional states eliciting expectations about upcoming referents (Intentional Account). That is, participants may have hypothesized that the robot attended to one object because it intended to mention it. On this account, congruent referential gaze elicits correct expectations about the utterance whereas incongruent gaze entails a revision in expectations upon hearing the actual reference, thereby slowing people.

In Experiment 3, we found evidence that the visual cue provided by robot gaze not only elicited a visual attention shift of the listener but indeed influences beliefs about the robot's referential intentions. Participants were asked to correct false robot utterances and were free to decide which (target or competitor) object they included in their correction sentence. Robot gaze modulated which object participants chose as referent in their correction statement, suggesting that their understanding about what the robot had originally intended to say was indeed affected by robot gaze. Thus, results from Experiment 3 supported the Intentional Account.

Importantly, in Experiments 1-3 we assumed that human-like synchronization of gaze and speech was required for people to be able to use gaze as a cue to the robots attentional and intentional states. The results from our previous experiments, however, allow no conclusion with respect to the relevance of temporal synchronization of gaze and speech for the observed effects. Moreover, we argue that the relevance of this synchronization could provide additional insights helping to illuminating the nature of gaze influence: If its effect on utterance comprehension exists only because speaker gaze happens to draws attention to the right object at the right time (Visual Account), changes in the temporal synchronization of gaze and speech should clearly affect the utility of gaze. In contrast, if people interpret robot gaze with respect to the robot's intentional states (Intentional Account), then synchronization is probably not critical. Understanding someone's (referential) intentions should persist and influence utterance comprehension as long as they seem relevant. A combination of the two accounts is also conceivable and, in fact, most probable. That is, *which* intention is inferred from the gaze cue depends on *when* the cue is produced relative to speech. The interpretation of robot gaze cues may, thus, be a function of temporal synchronization and linear order. Consequently, in Experiments 4 and 5, we manipulated these dimensions of gaze and speech alignment.

Recent findings from a study on the influence of indirect speaker gaze on utterance comprehension suggest that there is only limited flexibility in the requirement of temporal synchronization of such a gaze cue with speech, while maintaining its utility for the listener (Kreysa, 2009). Kreysa (2009) conducted several studies in which participants were shown a natural scene (photograph of a room) while listening to verbal descriptions of this scene previously given by another participant. Further, the speaker gaze of this participant was projected onto this scene such that listening participants saw an indirect gaze cue (a cursor which represented the speaker's original locus of fixation) while listening to the corresponding utterances. Listeners had to identify mentioned objects in the scene as soon as possible by clicking on them. To assess the importance of alignment between speaker gaze and speech, the indirect gaze cue was manipulated to occur one, two or five seconds before or after of the original cue. Kreysa (2009) found that as long as the shift with respect to the gaze cue's original occurrence was small (up to two seconds ahead or one second delayed of natural occurrence), the cue still facilitated the identification of referents. When the cue was more than two seconds ahead of its natural occurrence, or more than one second behind, people were slower to detect and click on the correct object. In fact, cues that were shifted by 5 seconds were no more beneficial than random cues shown in a baseline condition. Post-hoc tests between all conditions revealed a significant difference in click latencies between the natural condition and the baseline (random), but no difference between conditions with larger shifts and the random cursor. Kreysa concluded that natural timing of gaze and speech is optimal for listeners, i.e., most beneficial, and that a larger shift reduces the utility such that click latencies are increased and, in fact, similar to the random pattern exposure. Notably, the natural speaker gaze pattern she observed in her studies were typically in accordance with previous results on gaze and speech production, that is, fixations to an object peaked at approximately 800ms before onset of the object name (Kreysa, 2009, Chapter 4).

These results suggest that the effect of gaze on utterance comprehension is flexible only to a limited extent and not independent of the synchronization of gaze and the produced utterance. However, the gaze cues used in the discussed study are indirect cues and may not have the same status as direct perception of speaker gaze. Depending on how people perceive speaker gaze compared to the cursor used in Kreysa's study, we outline two different behaviors in response to speaker/robot gaze that is shifted considerably with respect to its original, 'natural' occurrence:

(a) Robot gaze is, similar to a gaze cursor, a *visual cue* that may (reflexively) direct attention and, thus, is only helpful for processing referring expressions when it occurs within a short time window around the spoken reference. A substantial shift of gaze relative to speech would result in longer response times than the original synchronization (and be no better than neutral gaze).

(b) Speakers' looks towards an object may be perceived as more *intentional* than a gaze cursor and as more robustly assigning relevance to the object in focus (similar to human gaze). Participants may persistently maintain and use this information when it seems relevant, leading to equal response time for shifted and synchronized gaze. Equally, non-congruent gaze cues may thus – even when shifted to precede the utterance – disrupt comprehension and cause slower response times.

Behavior (a) would provide support for the Visual Account (response level 2 in Section 1.4) whereas behavior (b) would again favor the Intentional Account (response level 3 in Section 1.4). Thus, Experiments 4 and 5 sought to investigate the importance of synchronization between gaze and speech for human-robot interaction, exploring the effects of shifted gaze cues on gaze-following, joint attention and its benefits for utterance comprehension.

Specifically, in Experiment 4 we investigated two kinds of synchronization. Firstly, we manipulated the temporal synchronization of gaze and speech by shifting gaze cues ahead of speech such that all gaze movements were completed before the robot utterance began (*preceding* condition). We compared participant behavior in response to *preceding* versus *synchronized* gaze in order to assess the significance of the temporal synchronization. Secondly, we manipulated the sequential order of mentioned references with respect to the order of gaze cues. That is, gaze cues that occurred in an order *reverse* to the order of mention were sequentially miss-aligned, which was compared to the *original* order of gaze and speech cues. This manipulation allowed us to explore people's ability to make use of a gaze cue which appears misplaced at the time of its occurrence but which is, nevertheless, relevant since it referred to a mentioned object.

Experiment 5 further investigated the facilitating/disruptive influence of the relative order of visual and linguistic references on utterance comprehension in greater detail. Using only synchronized stimuli, both original and reversed orders of referential gaze and speech cues were contrasted with neutral gaze. This way, we control for potential effects of the (sentential) order of mention in itself and investigate the actual benefit/disruption of coherent versus reversely synchronized gaze compared to neutral gaze. Thus, Experiments 4 and 5 together potentially provide additional insights into the general robustness of robot gaze for people's beliefs about referential intentions.

## 5.1. Experiment 4

In Experiments 2 and 3, we found that *congruent* robot gaze behavior, i.e. gaze to relevant objects one second prior to their mentioning, facilitates utterance comprehension compared to incongruent or neutral robot gaze. These experiments explored the effect of referential robot gaze with respect to its match with the uttered reference (referent identified by gaze cue was either mentioned or not). That this qualitative reference (mis-)match affected the speed and nature of referential processing shows, on the one hand, that people attend to an object that the robot seems to attend to and that people infer referential intentions from the robot gaze. On the other hand, this effect emphasizes the importance of *which* object the robot looks at but it does not allow specific conclusions with respect to the importance of *when* the robot looks at the according object.

In Experiments 1-3, we adopted the previously established timing of referential (speaker) gaze which precedes the onset of a referring noun by approximately one second. As noted above, Kreysa's results (2009) suggest that such precise timing is not essential while further showing that large deviations in synchronization reduce gaze benefits. In the present study, we consequently investigated whether referential robot gaze needs to be temporally aligned with speech (in the way human gaze is synchronized) in order to be beneficial, or whether robot gaze conveys referential intentions that have a long-lasting effect on utterance comprehension. To investigate this issue, we manipulated alignment in two ways. While robot gaze is always directed to the mentioned objects, i.e., never refers to irrelevant objects, we manipulated the factor Order of Mention (sequence of mentioned objects crossed with sequence of gazed at objects) which led to *original* (coherent) and *reverse* order of references. A second factor Synchronization manipulated the temporal shifts between gaze/visual reference relative to the corresponding linguistic references.

Recall that participants were shown a scene as given in Figure 5.1 and a corresponding unambiguous sentence such as "The cylinder is taller than the pink pyramid". To achieve the mentioned manipulations, we varied the sequence of referring expressions in the sentence: In the experimental condition "original order", the robot first mentioned the central object and then the peripheral object, while in "reverse order" the robot mentioned the peripheral object first, then the central object. Thus, the order of mention effectively manipulated the location of the referents for noun phrase one and two and constituted an experimental factor with two levels (original order: central-peripheral, reverse order: peripheral-central). The reversed order sentences were maintained valid by also reversing the comparative between two objects (see Table 5.1). Importantly, robot gaze was always directed first to the central object and subsequently to the peripheral object. Thus, a conflict arose in the sequence of these multi-modal references in the reverse condition as the sequence of gaze movements was in reverse order to the sequence of referring expressions. In the original order condition, both gaze cues and referring expressions had the same linear order.

The temporal synchronization of robot gaze and speech constituted the second factor with also two levels (*synchronous, preceding*). Either robot gaze was synchronized such that the robot fixated a referent one second prior to noun onset, or all gaze movements preceded the robot utterance completely such that the robot first gazed at the two referents and then uttered the sentence.

Notably, in all conditions the robot uttered a correct description and always gazed at the mentioned objects. Thus, the possible conflict in order of mention and order of gaze cues towards the referents may also be considered to reflect *temporal* congruency – not *absolute*, since the correct and mentioned objects are effectively being looked at. In terms of the previous manipulation of Congruency, conditions in Experiment 4 can be compared as follows: Original order of mention in combination with synchronized gaze equals *congruent* stimuli from Experiments 2 and 3, while reverse order of mention combined with synchronized gaze results in a reverse reference sequence (temporally

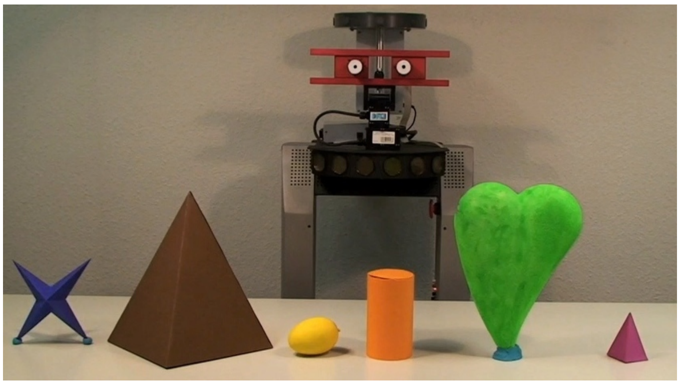


Figure 5.1.: Sample scene from Experiments, repeated here for convenience.

*incongruent*).

Uttered sentences in this experiment were similar to item sentences in Experiments 2 and 3 but were entirely unambiguous. That is, the adjective that previously disambiguated target or competitor in a relative clause after the referring noun was now changed to a prenominal adjective which already uniquely identified the referent. Sentences in this experiment further occurred in combination with identical scenes as in previous experiments. Figure 5.1 depicts a corresponding sample scene and Table 5.1 provides an example for each experimental condition. The examples are based on the sentence "The cylinder is taller than the pink pyramid" and show that robot gaze is always directed first to the cylinder and then to the pink pyramid. That is, the robot always looked first at the central object, then at the peripherally located object. We distinguished the location of the referents in this explicit way because it may affect visual attention as well as processing time. After all, items in the central visual field are typically more salient than those in the periphery (Mannan et al., 1996) and, thus, are potentially easier to access when linking the utterance to the scene. This may affect response time in particular when NP2 refers to the central object (reverse order) such that NP2 can be resolved faster and the utterance may be validated faster.

Table 5.1.: Order of Mention of mention, crossed with temporal synchronization of gaze and speech, results in these four different conditions. Angular brackets (< >) mark a gaze referent, quotation marks (") specify the linguistic referent. Effectively the sequences of (ling. and visual) references are provided for each condition.

| Robot Gaze | | | **Original** Order: *"The cylinder is taller than the pink pyramid."* | | | |
|---|---|---|---|---|---|---|
| **Synchronized** | | <cylinder> | "cylinder" | ... | <pyramid> | "pyramid" |
| **Preceding** | <cylinder> | <pyramid> | "cylinder" | ... | | "pyramid" |

| | | | **Reverse** Order: *"The pink pyramid is shorter than the cylinder."* | | | |
|---|---|---|---|---|---|---|
| **Synchronized** | | <cylinder> | "pyramid" | ... | <pyramid> | "cylinder" |
| **Preceding** | <cylinder> | <pyramid> | "pyramid" | ... | | "cylinder" |

### 5.1.1. Method

#### Participants

Thirty-two native speakers of German, again mainly students enrolled at Saarland University, took part in this study (6 males, 26 females). All participants reported normal or corrected-to-normal vision.

#### Materials

We manipulated two factors: Order of Mention (original, reverse) as well as Synchronization (synchronized, preceding). The four conditions of an item were created using one video stimulus (showing robot gaze to the central object and then to an object on the periphery of the table) and two different sentences with two different temporal onsets each (see Table 5.1). That is, each item appeared in four conditions. The scene and gaze movements shown in the video stimuli were identical to previous experiments. The two sentence types were of the form given in Table 5.1 and entirely unambiguous. That is, we removed the temporary referential ambiguity and shifted the LPoD to precede the noun by removing the relative sentence and substituting it by a prenominal adjective. This way, the relative importance of gaze with respect to unambiguous spoken references is examined to potentially corroborate the hypothesis that people generally follow and interpret robot gaze as constraining the domain of interpretation.

While the central object (*anchor* in previous experiments) was uniquely identified by naming its shape, objects on the periphery such as the pink pyramid required a disambiguating adjective in the noun phrase since there were two pyramids in the scene. The onset of the sentences varied according to the synchronization and was, in average, delayed by approximately 4300ms in the *preceding* condition compared to original occurrence. To counterbalance size, color and location of mentioned objects, we again created an equal set of four videos per item, each with a reversed comparative such that each peripheral object was a target referent in one set of stimuli and a potential competitor in another set (e.g. the brown pyramid in sample scene of Figure 5.1 became target when reversing the comparative). The approximate timing of a trial with synchronized gaze in reverse order is provided in Figure 5.2.

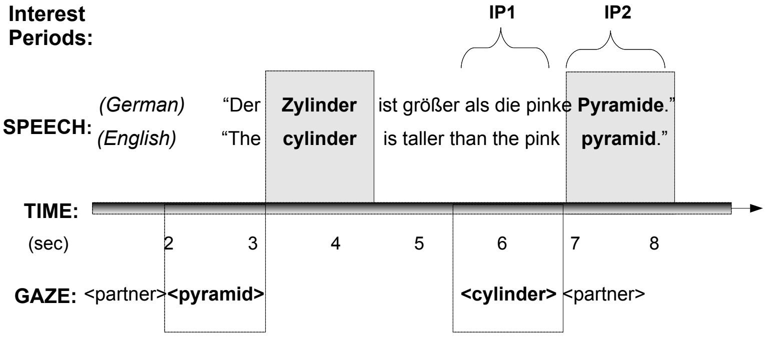


Figure 5.2.: The approximate timing of utterance-driven robot gaze, in a reverse-synchronized condition.

We constructed 36 filler videos and a total of 24 items resulting in 60 trials per list. Since all experimental items required a positive answer and participants were given a decision task, we introduced a bias towards negative answers in the fillers. 24 fillers contained incorrect statements which resulted in overall distribution of 40% incorrect trials per list. Otherwise, fillers were equally distributed across experimental conditions and varied only with respect to the comparisons they made and where mentioned objects were located in the scene.

Eight lists of stimuli were created, resulting from four experimental conditions and their counterbalanced versions. Each participant saw only one condition of an item and, in total, six stimuli in each condition. The order of item trials was randomized for each participant individually and items were always separated by at least one filler.

### Procedure

The task and procedure were identical to Experiments 1 and 2. That is, participants saw videos of the robot describing the scene and had to decide whether or not the robot's statements were correct. Participants' eye movements were again tracked during trials. The entire experiment lasted approximately 30 minutes.

### Analysis

The Interest Areas (IAs) in this experiment identified the central object, the peripheral (formerly target or competitor) object and the robot head. The "cylinder" from the example sentence above was the reference to the **central** and the "pink pyramid" was the reference to the **peripheral** object. For analyses we were particularly interested in the object mentioned in the final noun phrase henceforth called the NP2 referent, which was either the central or the peripheral object depending on the order of mention.

We segmented the speech stream into two Interest Periods (IPs). IP1 was defined as the 1000ms period ending at the onset of the second noun (in NP2). It contained the robot's fixation towards the target object as well as verbal content preceding the target noun. In this experiment, IP2 did not stretch from noun onset to offset but was defined as the 700ms period beginning with noun onset in NP2.

For inferential analyses, we considered inspections on the object referenced in NP2 as well as response time, recorded from NP2-noun onset to the moment of the button press. The analysis is otherwise identical to Experiments 1 and 2.

### Predictions

In Experiment 2, it was established that congruent robot gaze facilitates and incongruent gaze disrupts utterance comprehension. If this facilitation effect of gaze is a bottom-up process, i.e., it arises because the robot gaze cue draws attention to the *right* object at the *right* time (Visual Account), then a temporal shift of gaze cues relative to speech should diminish the benefit as well as the disruptive effect of gaze. However, the use of gaze could (simultaneously) be also a top-down process involving interpretation of gaze as an expression of attentional and intentional states of the robot, as indicated by the results from Experiment 3 (Intentional Account). If this is indeed the case, then the obtained information may be used to construct hypotheses about potential referents which persist throughout utterance processing until further constraining or contradictory information is obtained. Thus, the benefit of robot gaze is expected to be more robust and less sensitive to temporal shifts than the Visual Account predicts.

Specifically, we argue that the Intentional Account would not necessarily predict any reduction of gaze influence for shifted cues. Instead, it is plausible that intentions inferred from preceding gaze still facilitate or disrupt later reference resolution and lead to similar response times as synchronized gaze.

The Visual Account further predicts that the order of referential cues would affect their utility such that only original (coherent) order would facilitate comprehension. Thus, in the context of the Visual Account an interaction of the two factors Synchronization and Order of Mention is likely. That is, cues in original order would facilitate reference resolution compared to reversed order only when gaze is synchronized while in the preceding condition reverse and original order would elicit similar (slow) responses.

The Intentional Account also predicts an effect of Order of Mention. Previous results have shown that speaker gaze is directed at objects in the order of their mention (Griffin, 2001) such a listener is likely to also expect gaze and speech cues to occur in the same linear order. That is, originally ordered cues were predicted to cause shorter response times than reversed cues since expectations based on inferred referential intentions would be fulfilled in the order of their appearance.

Importantly, the Intentional Account predicts that a temporal shift does not affect the influence of robot gaze. That is, if reverse order of cues has a disruptive or less facilitating effect than cues in the same linear order, than this effect is expected to persist also in the preceding condition such that we predict no interaction between the factors Synchronization and Order of Mention.

### 5.1.2. Results and Discussion

#### Eye movements

Figures 5.3 and 5.4 show a plot of the eye-movement data for the whole duration of a trial. The initial two seconds of a trial were preview time, the robot head started moving at around 2,000ms after trial start. We started plotting after preview time and ended plotting just after the end of the robot utterance (9,500ms in the synchronized condition and 13,000ms in the preceding condition). We divided this large time window into 250ms-bins and computed fixation proportions for each IA (referent of NP1, referent of NP2 and robot head) within each bin. Fixations that did not fall within an IA were counted towards background fixations and are not included in the graph. The onsets of the nouns in NP1 and NP2 are marked in the graph as well as the occurrences of robot gaze and its target object. Robot gaze was always directed to the central object first, and then towards the peripheral object. Depending on the uttered sentence (order of mention), the fixated object was mentioned in NP1 (marked as "np1 gaze") or in NP2 ("np2 gaze"). Final bins in each graph may be disregarded as they span the end of the average response time, i.e., contain sparse data and are unlikely to reflect general patterns.

As noted above, that the manipulation of Order of Mention coincided with a difference in location of the NP2 referent. That is, in original order, the NP2 referent is located in the periphery of the table (pink pyramid), while in reverse order it is located in the center of the scene (cylinder), as depicted in Figure 5.1.

Shown in the time graph are fixations on the NP1 and NP2 referents as well as on the robot head. It is clearly visible that people fixated the robot head more frequently than the objects on the table until the robot started speaking. Moreover, during robot gaze movements people followed the gaze to the respective objects (first central, then peripheral object), notably while fixating the robot head rarely. In the condition original-synchronized, the object referred to by robot gaze and subsequently by NP1 was identical and participants smoothly continued to fixate the according IA. Similarly, people followed robot gaze to the object which was then mentioned in NP2. In condition original-preceding, participants similarly followed robot gaze to the NP1 referent and the NP2 referent before looking back to the NP1 referent (central object) when the robot starts speaking. The fixation pattern throughout the robot utterance is remarkably similar to the pattern observed in the original-synchronized condition.

In the lower graphs depicting reverse-order, people fixated what the robot initially fixated (which is again the central object, but now mentioned only in NP2, hence 'np2-gaze'). Then participants redirect visual attention towards the mentioned object (NP1 referent).

Notably, in all conditions participants frequently looked at the NP2 referent prior to its mention, irrespective of the order of mention, gaze direction and gaze synchronization. It is not clear whether participants indeed used even reverse gaze such that they anticipated the NP2 referent in all conditions or whether this fixation pattern also re-

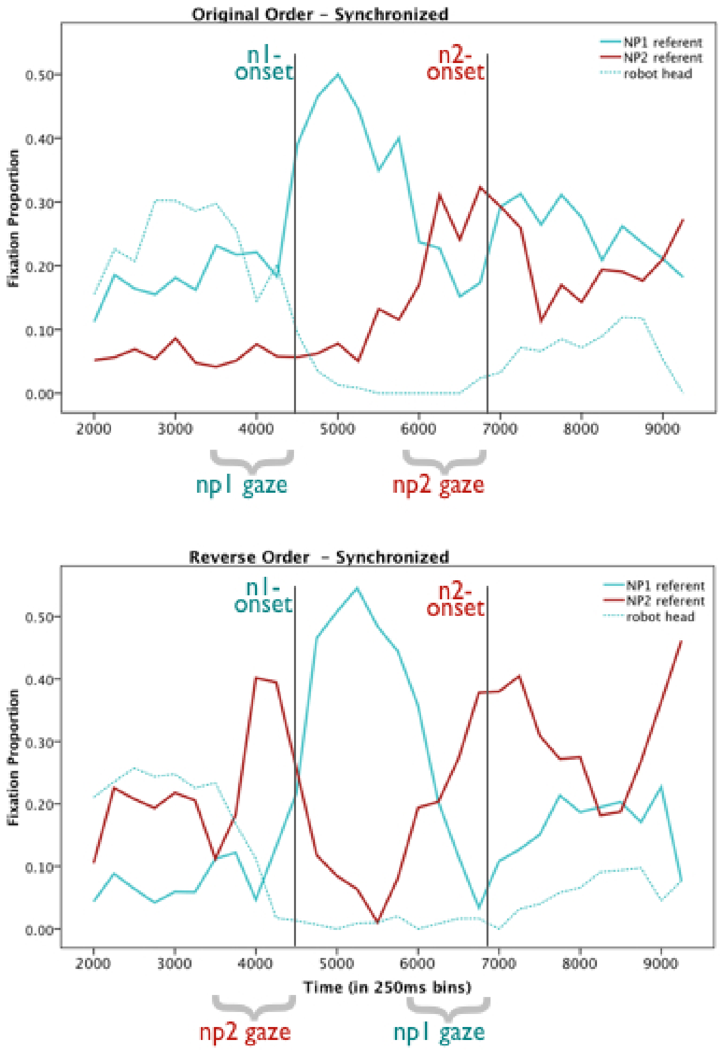


Figure 5.3.: Time graph for **synchronized** robot gaze. Note that, robot gaze is directed first to central then to peripheral object. Depending on order of mention, central object is mentioned in NP1 and peripheral object in NP2 or vice versa, hence *np1 gaze* or *np2 gaze*. IP1 ends and IP2 begins with noun onset in NP2.

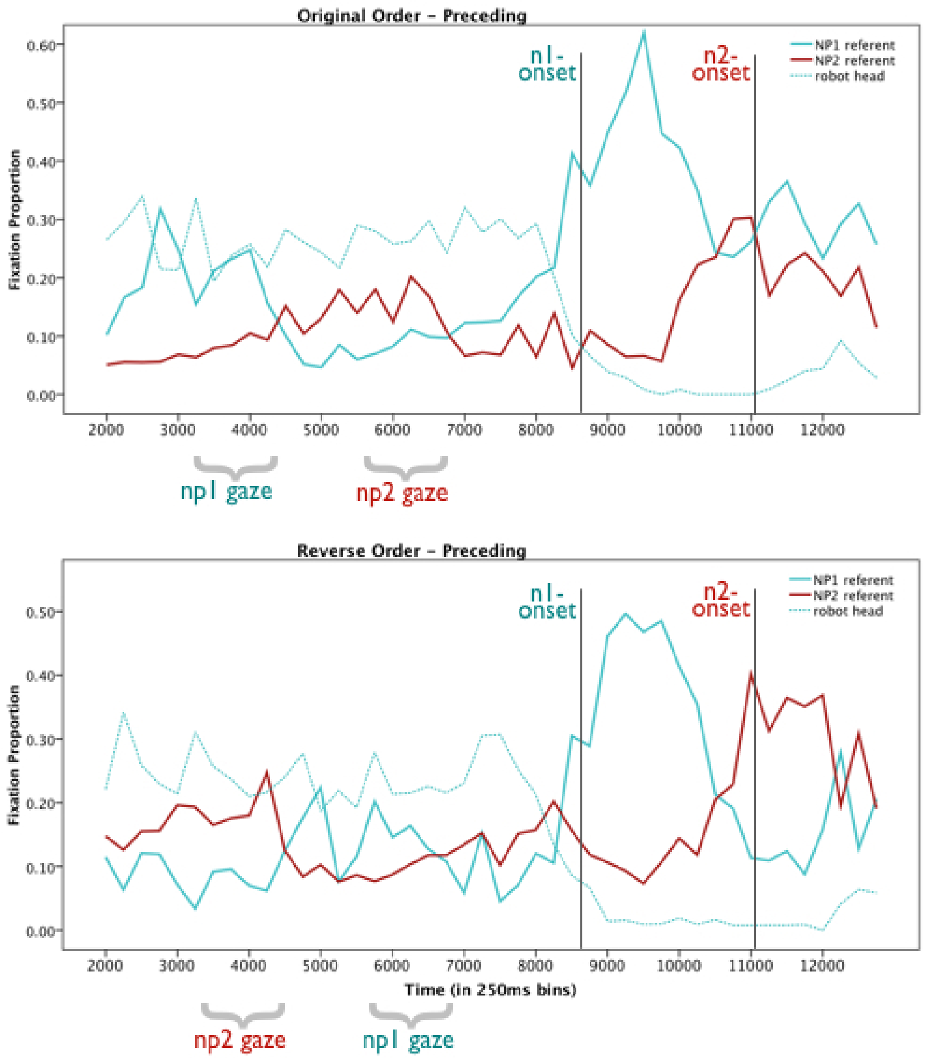


Figure 5.4.: Time graph for **preceding** robot gaze. Note that, robot gaze is directed first to central then to peripheral object. Depending on order of mention, central object is mentioned in NP1 and peripheral object in NP2 or vice versa, hence *np1 gaze* or *np2 gaze*. IP1 ends and IP2 begins with noun onset in NP2.

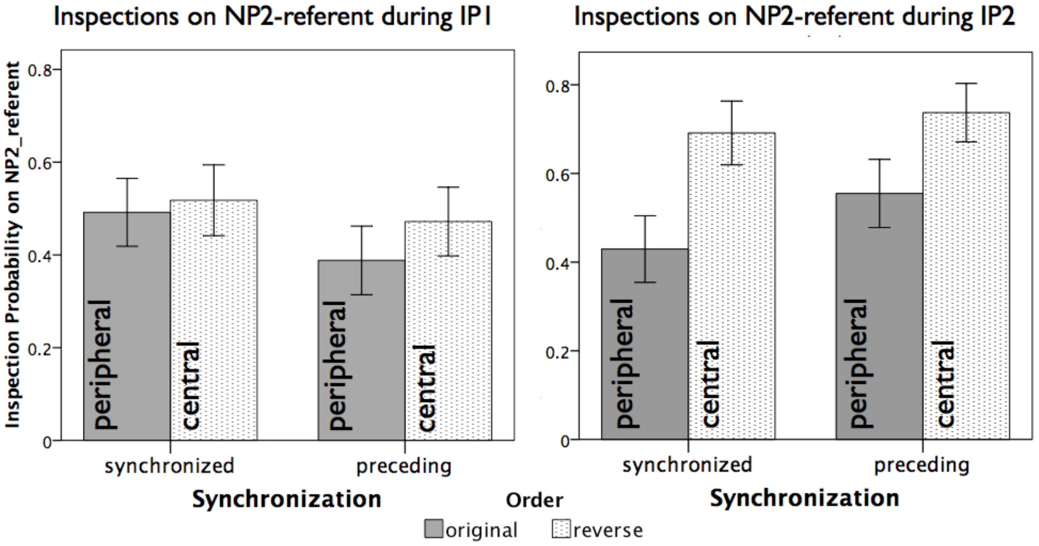


Figure 5.5.: Mean inspection probabilities per condition for IP1 (left graph) and IP2 (right graph). IP1 is the 1,000ms time window preceding the noun onset in NP2. IP2 is defined as the 700ms time window starting at noun onset in NP2. Further, bars are labeled with respect to the location of the NP2 referent in that condition.

flects other processes. Consider an example: <Robot looks at (central) cylinder> "The pink pyramid is taller than <robot looks at (peripheral) pyramid> the cylinder." When the robot looks at the pink pyramid, participants have already heard "The pink pyramid is taller than" and hypothesize that the robot is not going to mention the pink pyramid again. Instead they may remember that the robot initially looked at the cylinder and use this piece of information to predict the NP2 referent. The time graph suggests that this may well be happening in the *preceding* condition. We consider a second example (original, preceding): <Robot looks at (central) cylinder> <robot looks at (peripheral) pyramid> "The cylinder is taller than the pink pyramid." By the time participants hear "taller than the", they already fixate the (peripherally located) pink pyramid which suggests that they have inferred some referential intention from the robot's prior gaze movements, now predicting the pyramid to be the NP2 referent. While these time graphs depict averaged eye movements which reflect tendencies for visual attention direction, inferential statistics of both eye movements and response time data will reveal, for instance, whether people indeed map robot gaze to the utterance quickly enough to facilitate comprehension or whether an incongruent sequence of references interrupts comprehension.

Mean inspection probabilities for the NP2 referent are depicted in Figure 5.5. Results from inferential statistics on inspection data for IP1 and IP2 are given in Table 5.2. In IP1, both model reduction ($\chi^2(1) = 4.03; p < 0.05$) as well as the predictor's coefficient reveal a main effect of Synchronization on inspections on the NP2 referent. As the negative coefficient suggests, people inspected the NP2 referent with lower probability when robot gaze preceded the utterance than when it was synchronized. Moreover, we did not observe a main effect for Order of Mention, i.e., people inspected the NP2 referent equally often irrespective of where this referent was located (centrally, peripherally) or whether the robot concurrently fixated this object. This finding indicates that people may use the visual information provided by robot gaze cues, across conditions, to at least *visually* anticipate NP2.

In IP2, we observed a slightly different inspection pattern. In this IP, Order of Mention had a main effect on inspection probability ($\chi^2(1) = 35.67; p < 0.001$) such that participants inspected the mentioned object significantly more often in the reverse condition than in the original, coherent order of reference. That is, when the robot had previously fixated the peripheral object in IP1 and then mentioned the other, central object in IP2 (i.e. <gaze to central cylinder>"The pink pyramid is taller than <gaze to peripheral pyramid> the cylinder.") participants were more likely to inspect the mentioned object.

There are two possible explanations these high probabilities of inspecting the NP2 referent in reverse order: Either participants inspected this central object more often because it was more salient due to its central location, predicting easy and quick reference resolution. Alternatively, the increased inspections on the NP2 referent in reverse condition reflect difficulty to resolve the reference as it includes conflicting information (gaze identified the pyramid while the mentioned noun referred to the cylinder) – in which case slower response times would be expected. The latter explanation seems to conflict with the assumption that people indeed predicted NP2 from gaze cues in all conditions. Upon further consideration, this is not a real conflict, however. Even though visual attention is directed to the correct object, mapping the gaze cue to resolve the reference and integrating this piece of information into the utterance comprehension process may result in a greater cognitive load compared to the coherent sequence of multi-modal references. The response time results will reveal which of the two explanations is more likely.

Table 5.2.: Model fitted to inspection data on object mentioned in second noun phrase, in IP1 and IP2.

| | Predictor | Coefficient | *SE* | Wald Z | p |
|---|---|---|---|---|---|
| IP1 | (Intercept) | -0.0424 | 0.1758 | -0.241 | 0.809 |
| | Order - reverse | 0.1122 | 0.2165 | 0.518 | 0.604 |
| | Synchronization - preceding | -0.4372 | 0.2187 | -1.999 | <0.05 |
| | reverse:preceding | 0.2462 | 0.3089 | 0.797 | 0.426 |
| IP2 | (Intercept) | -0.2917 | 0.1680 | -1.737 | 0.083 |
| | Order - reverse | 1.1128 | 0.2314 | 4.808 | <0.001 |
| | Synchronization - preceding | 0.5056 | 0.2218 | 2.280 | <0.05 |
| | reverse:preceding | -0.2773 | 0.3293 | -0.842 | 0.399 |

$$Model : NP2referent \sim OrderOfMention * Synchronization$$
$$+(1|subject) + (1|item), family = binomial(link = "logit")$$

It should be pointed out, however, that the reversal of item sentences produced asymmetric conditions: NP2 contained a disambiguating adjective in the original order-of-mention condition while this adjective was not present in the reverse order-of-mention condition. The reason for this is that a color adjective for the central object would have been redundant since the noun uniquely identified the object. This imbalance resulted in a confound of the manipulation of Order of Mention with the presence of an adjective in NP2. Both variations make the same predictions with respect to response times. Both a coherent sequence of multi-modal references (original order) as well as an additional adjective in NP2 were predicted to facilitate reference resolution and result in shorter response time. A closer look at participants' eye movements may help to identify which experimental manipulation accounts for potential response time effects. In order to incorporate the possible influence of the adjective in IP1 – where we observed gaze-mediated and/or anticipatory eye movements to the object about to be mentioned in IP2 – we shifted both IPs to 200ms later. That is, IP1-shifted was slightly shortened and stretched from 500ms prior to noun onset to 200ms after noun onset and

Table 5.3.: Model fitted to inspection data on object mentioned in second noun phrase, in **shifted** IPs.

| | Predictor | Coefficient | *SE* | Wald Z | p |
|---|---|---|---|---|---|
| IP1-shifted | (Intercept) | 0.3112 | 0.1724 | 1.806 | 0.071 |
| | Order - reverse | -0.0912 | 0.2219 | -0.411 | 0.681 |
| | Synchronization - preceding | -0.4659 | 0.2266 | -2.056 | <0.05 |
| | reverse:preceding | -0.0412 | 0.3181 | -0.129 | 0.897 |
| IP2-shifted | (Intercept) | -0.6989 | 0.1841 | -3.797 | <0.001 |
| | Order - reverse | 1.4774 | 0.2392 | 6.176 | <0.001 |
| | Synchronization - preceding | 0.2141 | 0.2330 | 0.919 | 0.358 |
| | reverse:preceding | 0.0245 | 0.3356 | 0.073 | 0.942 |

$$Modelfinal : NP2referent \sim OrderOfMention * Synchronization + (1|subject) \\ +(1|item), family = binomial(link = "logit")$$

IP2-shifted was defined as the subsequent 700ms period. Thus, IP1-shifted accommodated the time needed to process the adjective such that potentially resulting anticipatory eye-movement effects are captured in this time window. If the NP2-adjective indeed helped participants to anticipate the object referenced by NP2 (potentially resulting in short response time), then an effect of Order of Mention would be predicted in IP1-shifted. That is, already in IP1-shifted, we would expect more inspections on the NP2 referent in the original order of mention, compared to the reverse order condition.

As shown in Table 5.3, there is no main effect of Order of Mention in IP1-shifted, suggesting that the adjective in NP2, in reverse order, does not affect anticipation of the referent significantly. Instead, the Synchronization effect already reported from the initial IP1 is still present, suggesting that people indeed followed gaze and attended more closely to the according objects during actual robot gaze movements.

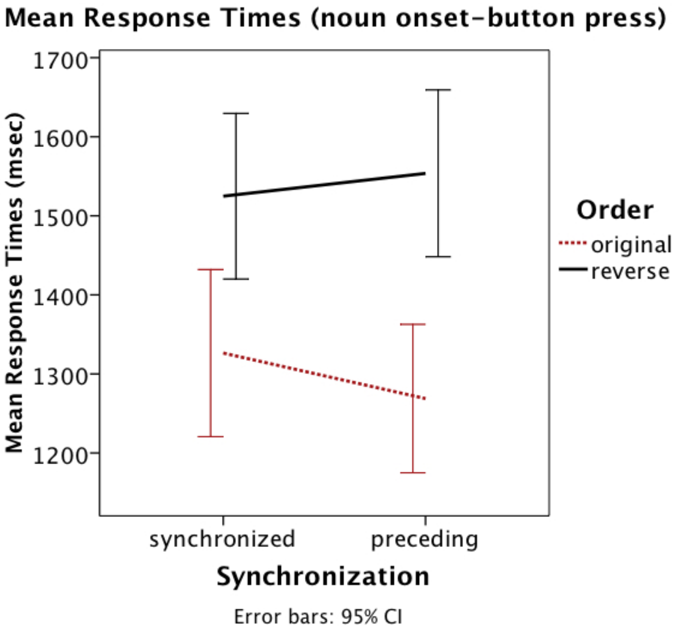


Figure 5.6.: Average response times for all four conditions.

### Response Time

Model reduction shows that Synchronization had no effect on response times. That is, participants were equally fast to determine the validity of the robot statement in synchronized and preceding conditions. Since no interaction between the two factors Synchronization and Order of Mention was observed, we excluded Synchronization as a predictor from our model. In addition, model reduction reveals a main effect of Order of Mention ($\chi^2(1) = 45.19; p < 0.001$, see also Figure 5.6 for averages).

The result suggests that the temporal shift of robot gaze from synchronized to preceding the utterance did not affect the utility of the gaze cues, but the order of the cues did. This, however, contradicts the predictions derived from a purely Visual Account which suggested that it was critical *when* robot gaze drew attention to an object. Both manipulations, however, made robot gaze direct people's attention to relevant objects at non-synchronized points in time – and always prior to the last referring expression (NP2) – such that a difference in how these factors affect utterance processing can only be explained in terms of an Intentional Account: The precise temporal synchronization is not crucial for people to interpret and use robot gaze as a cue to the robot's intentions. The inferred (referential) intentions, however, are expected to be fulfilled in the *same order* as they were indicated by the robot's gaze. This is not surprising since the order of gaze cues reflects the speaker's intentions regarding order of mention (Griffin and Bock, 2000; Griffin, 2001). Thus, people seem to expect that the inferred referential intentions be realized in the corresponding order. If this expectation is not met, gaze cues – even when identifying mentioned objects – disrupt comprehension.

One might be concerned that the response time findings for Order of Mention resulted from the sentence order itself in the event that reverse order of mention is generally more difficult to process due to, for instance, a certain visual search effort during reference resolution. However, visual search involved in resolving NP2 is in fact less extensive in this experimental condition since the central object is mentioned in NP2 and this centrally located object is arguably most salient and easiest to find. If the peripheral object location influences the effort needed to resolve the referent, then a potential difficulty should occur at the beginning of the sentence (NP1, peripheral object) and would have most likely been resolved by the end of the sentence. This suggests that, if any difference is expected at all, reverse order of mention should result in *reduced* response times. Since this is not the case, we suggest that the increased response time reflect the conflict in order of mention and order of gaze cues. This conflict seems to be caused by gaze cues that elicit expectations about a certain sequence of referring expressions which are not met, even in the case when robot gaze precedes the utterance. In contrast, the close temporal synchronization of gaze and speech seems to be generally less important, indicating that – regardless of timing – robot gaze evokes expectations about the robot's attentional (and possibly intentional) states during interaction.

### Summary

In this experiment, we have manipulated the alignment of referential gaze and speech cues in order to examine the flexibility of this alignment while maintaining the benefit of gaze for utterance comprehension. More precisely, we considered two kinds of alignment: Firstly, we manipulated the temporal synchronization of gaze and speech: gaze cue were either *preceding* (all gaze cues were shifted such that they preceded the robot utterance) or *synchronized* (gaze and speech cues were produced concurrently, in reverse order even in an overlapping manner). Secondly, we manipulated the order of referring expressions and referential gaze cues. That is, gaze and speech cues were either in *original* (and coherent) order or in *reverse* order to each other.

We found evidence that the precise temporal synchronization is not critical for the utility of robot gaze. A substantial temporal shift of roughly 4.3 seconds of the gaze cues relative to their 'natural' occurrence (preceding condition) caused the same effects as synchronized gaze. That is, when gaze was in the same linear order as speech it similarly facilitated comprehension in the synchronized (which has been shown to facilitate comprehension in Experiment 2) and the preceding condition. When gaze occurred in a reverse order to speech, this had a similarly slowing effect on response times in both synchronized and preceding conditions.

This result suggests, that people follow and use robot gaze for utterance comprehension even after a considerable period of time. Notably, a purely attentional explanation for comprehension facilitation (Visual Account) suggesting that gaze happens to draw attention to an object which is then mentioned, would have predicted that the utility of gaze is affected by a substantial temporal shift (e.g., as shown by Kreysa, 2009). The Intentional Account, in contrast, is consistent with the notion that people infer intentional states from robot gaze and therefore predicted that people maintain and use the provided information when it seems appropriate (or until outdated). Thus, the results on temporal synchronization effects provide more evidence in favor of the Intentional Account and, thus, also in favor of response level 3 introduced in Section 1.4.

Interestingly, we further observed that the order of referential cues significantly affected the benefit of robot gaze for utterance comprehension. That is, while the precise temporal synchronization of gaze and speech was not crucial for the utility of robot gaze, the relative ordering of cues did affect response times. Obviously, the inferred (referential) intentions were maintained over several seconds but were also expected to be fulfilled in the *same order* as they were indicated by the robot's gaze.

The simultaneous absence of a main effect of Synchronization and presence of a main effect of Order of Mention is evidence for the importance of when a referential gaze cues occurs *relative* to the according referring expression. However, two further questions arise from this study which need to be addressed in the following experiment. Despite the inspection analyses suggesting that an imbalanced use of a prenominal adjective did not affect referent anticipation, we cannot conclude that it had no effect on response times. Thus, in Experiment 5 we remove this confound and further examine the effect of Order of Mention. Additionally, the results from this experiment do not reveal whether cues in reverse order actually disrupt comprehension or whether they are simply not as beneficial as cues in original (and coherent) order. This issue is examined in Experiment 5 by contrasting original and reversed gaze and speech cues with neutral gaze.

## 5.2. Experiment 5

Results from Experiment 4 demonstrated that a close temporal coupling is not essential for robot gaze to influence utterance comprehension suggesting that gaze is interpreted as a cue to the robot's referential intentions rather than providing a purely visual cue. In contrast to a related study conducted by Kreysa et al. (2009), the robot's gaze involved very few saccades over the course of a trial while it seemed to provide a very reliable cue in the sense that people used it even when it preceded the whole utterance. Since in Experiment 4 no comparisons between manipulated gaze and neutral gaze conditions were made, it was left open whether reverse order was actually disrupting or simply not facilitating comprehension (in contrast to original order). Experiment 5, thus, investigated the beneficial or disruptive effects of reversed robot gaze (and speech) cues compared to neutral gaze with the aims of providing insights into whether or not people are disrupted by incorrect gaze order or, in contrast, even use it to resolve referring expressions faster than when only neutral gaze is available. Consequently, results from this study complement our previous results, especially those from Experiments 2 and 4 which provided evidence for a disruptive effect of incongruent gaze. While incongruent behavior in Experiment 2 was caused by a referential mismatch such that one modality referred to an irrelevant object, in the current study (as in Experiment 4) both modal cues referred to relevant (mentioned and correct) objects but in a different order.

To investigate these issues, we again manipulated Order of Mention (original, reverse) and Synchronization (synchronized, neutral) by contrasting synchronized gaze and speech with neutral gaze, rather than preceding gaze. Since the temporal shift from synchronized to preceding condition in Experiment 4 did not substantially affect people's behavior, we did not include preceding gaze as a condition and instead considered neutral gaze. Firstly, combining original and reverse order of mention with neutral gaze allowed us to investigate whether the order of spoken references alone affects utterance comprehension. And secondly, the comparison between neutral gaze and synchronized gaze evaluated the facilitating/disruptive effect of originally or reversely synchronized gaze cues with respect to a neutral baseline.

The use of synchronized gaze crossed with Order of Mention essentially replicated the level synchronized-original and synchronized-reverse used in Experiment 4. That is, synchronized gaze and reversed order of spoken references resulted in reverse order of referential (visual and linguistic) cues, whereas synchronized gaze and original order of spoken references resulted in a congruent sequence of referential cues.

Table 5.4.: Order of Mention, crossed with Synchronization of gaze and speech, results in four different conditions. Angular brackets (< >) mark a gaze referent, quotation marks (") specify the linguistic referent. Effectively the sequences of (ling. and visual) references are provided for each condition.

| | **Original** Order: | | | | |
|---|---|---|---|---|---|
| Robot Gaze | *"The orange cylinder is taller than the pink pyramid."* | | | | |
| **Synchronized** | <cylinder> | "cylinder" | ... | <pyramid> | "pyramid" |
| **Neutral** | <> | "cylinder" | ... | <> | "pyramid" |

| | **Reverse** Order: | | | | |
|---|---|---|---|---|---|
| | *"The pink pyramid is shorter than the orange cylinder."* | | | | |
| **Synchronized** | <cylinder> | "pyramid" | ... | <pyramid> | "cylinder" |
| **Neutral** | <> | "pyramid" | ... | <> | "cylinder" |

The experimental conditions used in this experiment are described below using the sample sentence "The orange cylinder is taller than the pink pyramid." Note that robot gaze was always directed first to the central object, here the cylinder, and then to the peripherally located object, the pink pyramid. The sentences were, in contrast to Experiment 4, fully symmetric such that each noun phrase contained an adjective. This symmetry made sure that the final referring expression (NP2) was similar across all conditions such that there was no confound of condition and prenominal adjective occurrence. Otherwise, sentences and scenes were similar to the material used in the previous experiments.

### 5.2.1. Method

#### Participants and Procedure

Thirty-two native speakers of German, mostly students enrolled at Saarland University, took part in this study (11 males, 21 females). All reported normal or corrected-to-normal vision. Task and Procedure were identical to Experiment 4.

### Materials

A set of 20 items was used. The four conditions of one item were created using two different video stimuli and two different sentences. The videos varied according to Gaze Synchronization, that is, the *synchronized* condition showed robot gaze to the central object and then to an object in the periphery of the table, while the *neutral* condition showed the robot's initial glance down at the scene before looking up and beginning to speak.

The spoken sentences varied according to Order of Mention such that either the central object was mentioned followed by the peripheral object (*original*) or vice versa (*reverse*). A combination of both video stimuli and both sentence versions resulted in the four conditions depicted in Table 5.4. We further constructed 32 fillers for these items. Since all experimental items required a positive answer and the task was a decision task, we again introduced a bias towards negative answers in the fillers. Thus, 24 fillers (75%) contained incorrect statements which resulted in an overall distribution of 46% false trials. Because we had two conditions showing neutral gaze compared to only one showing original order cues and one showing reversed ordering of referential cues, fillers were distributed across conditions such that an equal number of trials had neutral gaze (18, in both original and reverse order), original-synchronized (17) and reverse-synchronized gaze (17). The two experimental conditions showing neutral robot gaze provided a means to assess the influence of order of mention as such. If there was no specific advantage or disadvantage related to Order of Mention, facilitation/disruption effects can be assigned to the relative order of referential cues (also retrospectively for effects found in Experiment 4).

### Analysis

The IAs in this experiment were identical to Experiment 4 and contained the central object, the peripherally located object and the robot head. Since robot gaze was synchronized with speech such that a robot gaze shift towards an object ended on a referent one second prior to noun onset, IP1 was defined to *begin* 1,000ms prior to noun onset in NP2. However, in this experiment IP1 does not stretch to noun onset but already ends with *adjective onset*. Thus, IP1 has no fixed duration but an average length of 600ms. This shortening of IP1 was done to incorporate the fact that the adjective preceding the noun unambiguously identified the referent. Consequently, IP2 was defined to stretch from adjective onset to 700ms after noun onset and had a mean duration of

1,100ms. Defining IP1 and IP2 in this way made it possible to distinguish once again between gaze-mediated inspections in IP1 and utterance-mediated inspections in IP2. The approximate timing of a trial with synchronized gaze in reverse order is visualized in Figure 5.7.

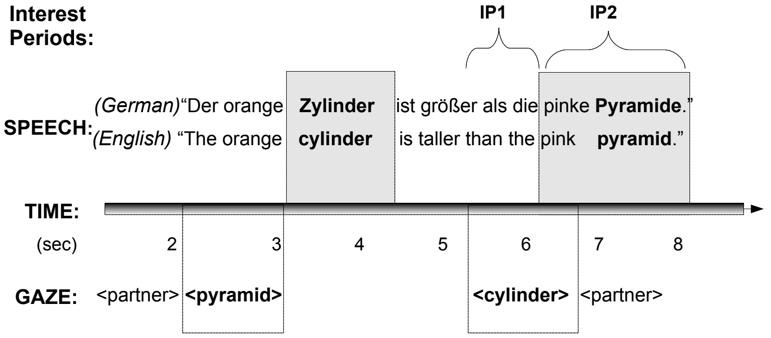


Figure 5.7.: The approximate timing of utterance-driven robot gaze, in a reverse-synchronized condition.

For inferential analyses, we considered inspections on the object referenced in NP2 as well as response time, recorded from NP2-*adjective onset* to the moment of the button press. The analysis was otherwise identical to previous experiments.

### Predictions

In the synchronized condition, we expected to replicate the findings from Experiment 4. That is, the order of produced gaze and speech was predicted to be relevant such that listeners only benefited from gaze if it was aligned to speech in the same linear order. We expected people to be again slower in validating the robot's utterance when the referential order of gaze and speech differed (i.e., was reversed).

The neutral gaze conditions further establish a baseline to determine the facilitation/disruption effects of synchronized gaze. Specifically, the comparison between neutral and synchronized conditions was predicted to reveal whether reversely coordinated gaze did simply not facilitate comprehension in the same way that an original order of cues did, or whether this, in fact, disrupted comprehension.

### 5.2.2. Results and Discussion

#### Eye movements

Fixation proportions on the NP1 and NP2 referents as well as on the robot head are plotted in Figures 5.8 and 5.9. Noun onsets are marked as distinct events in the unfolding speech stream. However, IP1 offset (= IP2 onset) is approximately 400ms prior to the second noun onset, at adjective onset. The plot clearly shows that, similar to Experiment 4, participants followed robot gaze and speech and inspected the looked at, and then mentioned, objects.

*Neutral Gaze:* Interestingly, the two neutral conditions reveal a fundamental difference of the Order of Mention. Contrary to what we predicted, namely that the order itself would have no significant influence, we observed considerable differences in participants' eye movements. The plot showing condition original-neutral indicates that the central object was initially the most salient object and, thus, was fixated before mentioning. This was probably due to the fact that objects centrally located in the visual field were more salient than others and, maybe more importantly the robot produced an initial glance down at the scene – and back up – before beginning to speak. This gaze movement was inserted to add some robot movement to neutral trial videos instead of presenting a more or less still frame, thereby making these conditions appear equally 'live'. However, the central object may have become even more prominent through this initial glance which may explain why people, in original order, fixated NP1 referent (the central object) already before noun (and even adjective) onset. Similarly, in reverse order, the NP2 referent (again the central object) was fixated already before noun onset. With regard to this preference for the central object, two inspection patterns become apparent: While original order seemed to be an advantage for processing NP1, NP2 was hardly anticipated and people fixated its referent only after noun onset. Reverse order of mentioning, on the other hand, seemed to facilitate anticipation and processing of NP2. This is reasonable as the peripherally located object had been mentioned in NP1 such that people fixated this object during mentioning and then re-directed their attention to the center of the scene (possibly preparing for further visual search) where the NP2 referent is located.

*Synchronized Gaze:* The fixations pattern shown in the plot for condition reverse-synchronized (incongruent order of referential cues) is similar to that observed for reverse-neutral. This is somewhat surprising, since people did not seem to follow the reversed gaze cues. Instead, people seemed to anticipate the NP2 referent (central ob-

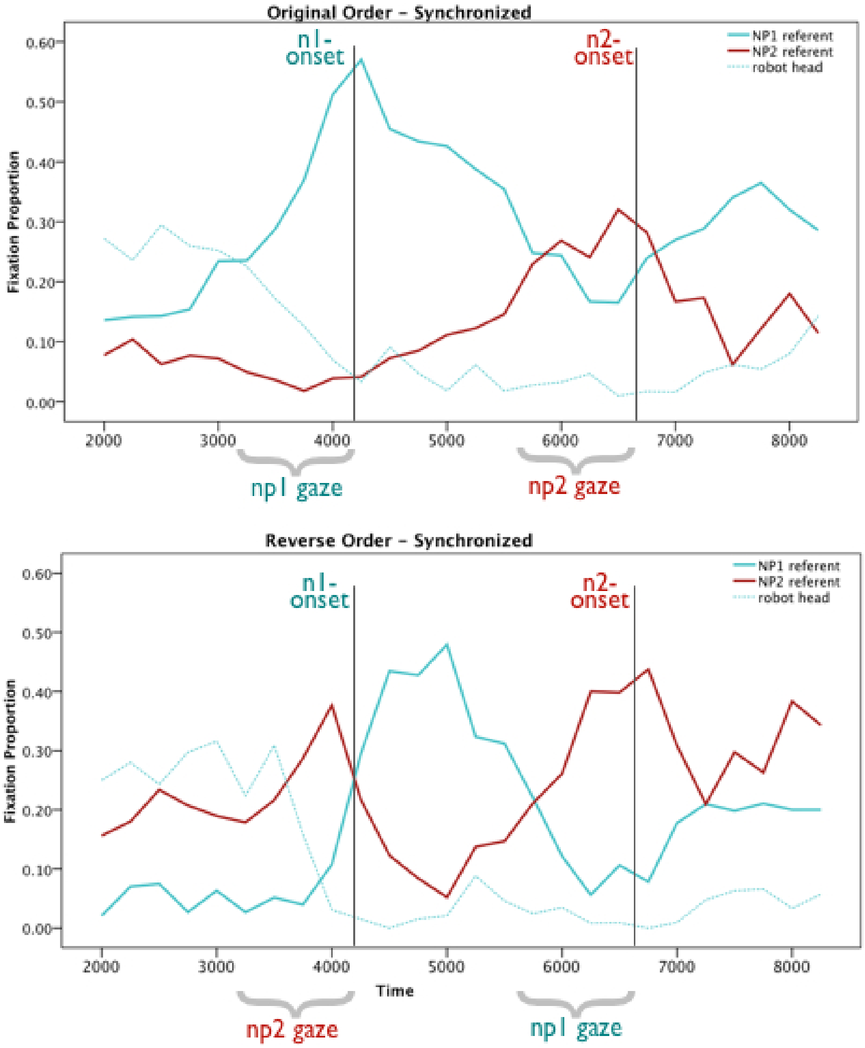


Figure 5.8.: Time graph for **synchronized** robot gaze. Note that, robot gaze is directed first to central then to peripheral object. Depending on Order of Mention, central object is mentioned in NP1 and peripheral object in NP2 or vice versa, hence *np1 gaze* or *np2 gaze*.

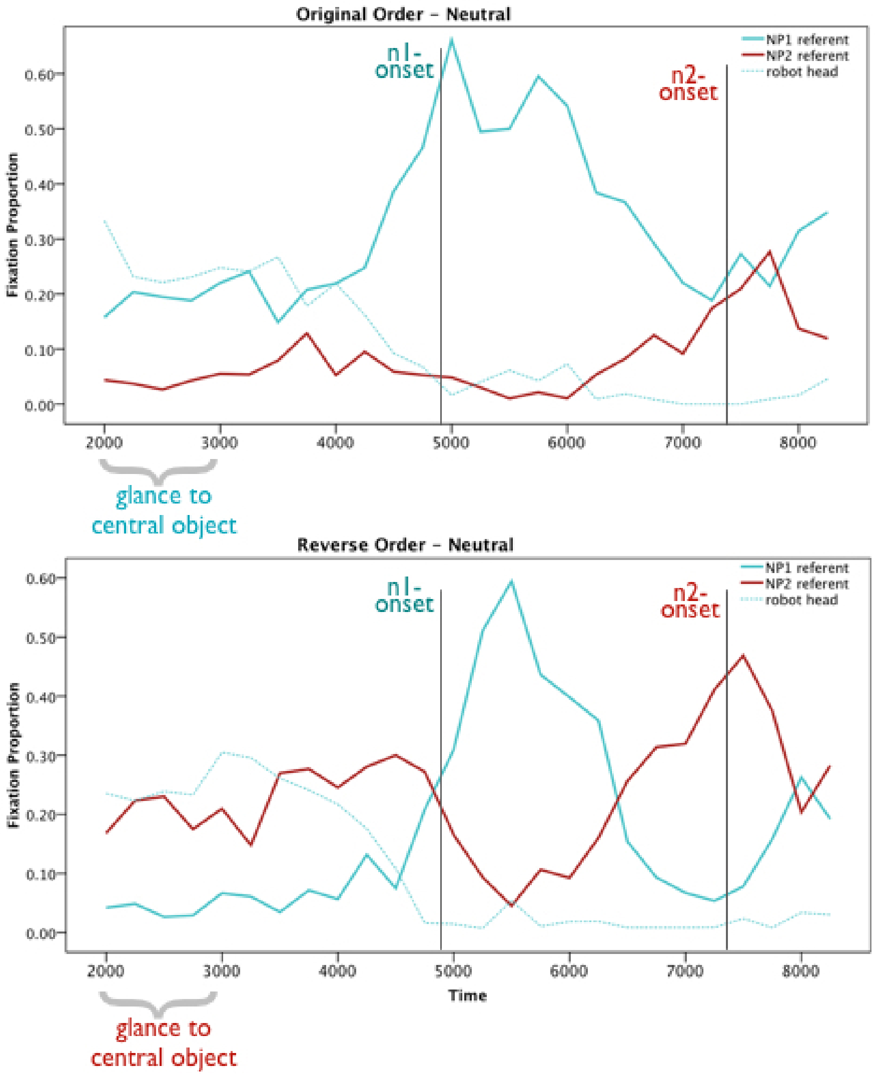


Figure 5.9.: Time graph for **neutral** robot gaze. The initial glance down onto the scene may direct people's visual attention towards the centrally located object. Depending on Order of Mention, this object is mentioned first or second.

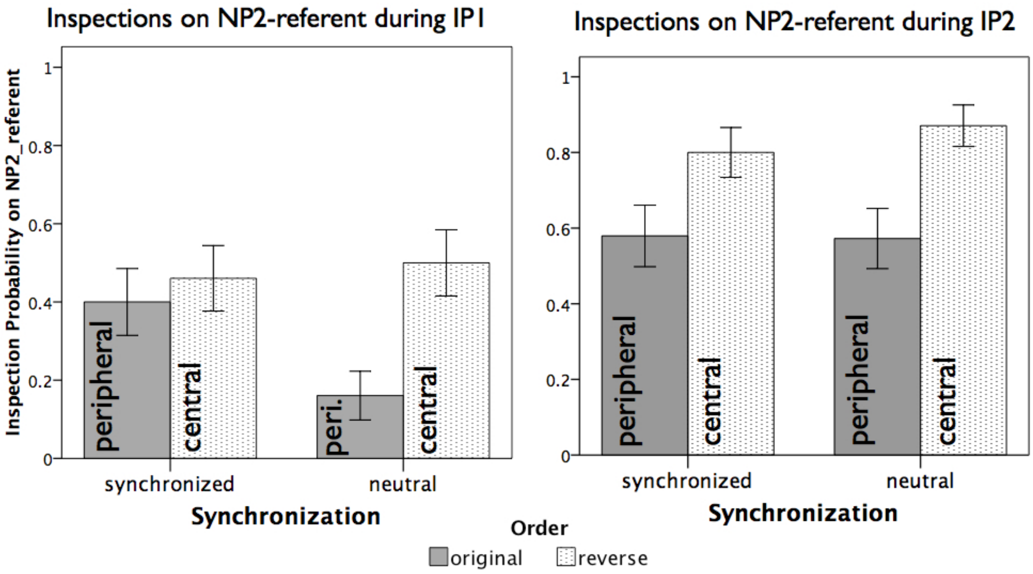


Figure 5.10.: Mean inspection probabilities per condition for IP1 (left graph) and IP2 (right graph). IP1 is the 600ms time window preceding the adjective onset in NP2. IP2 stretches from adjective onset in NP2 to noun offset in NP2. Further, bars are labeled with respect to the location of the NP2 referent in that condition.

ject) in the second gaze period – prior to NP2 – even though robot gaze was directed towards the peripheral object. However, at this point the peripheral object had already been mentioned in NP1 which may have caused people not to inspect it any further. The plot of condition original-synchronized (coherent order of referential cues) suggests that, similar to original-neutral, people visually anticipated the NP1 referent and continued to fixate it during mention. In contrast to the neutral condition, gaze towards the NP2 referent – preceding its mention – was then available to participants who followed the robot's gaze, using it to anticipate NP2. Note, that the stimuli in both conditions with temporally *synchronized* gaze (original and reverse order) were similar to the two conditions that showed synchronized gaze in Experiment 4. Accordingly, the observed fixation pattern for these conditions is also similar in both experiments.

Inferential statistics mainly confirm the observations from the time graphs. Mean probabilities for inspecting the NP2 referent are given in Figure 5.10 and results from inferential analyses are provided for both IPs in Table 5.5. In IP1, both Order of Mention and Synchronization had main effects on inspection behavior (Order: $\chi^2(1) =$

Table 5.5.: Model fitted to inspection data on object mentioned in second noun phrase, in IP1 and IP2.

| | Predictor | Coefficient | *SE* | Wald Z | p |
|---|---|---|---|---|---|
| IP1 | (Intercept) | -0.4402 | 0.1954 | -2.252 | 0.024 |
| | Order - reverse | 0.2733 | 0.2493 | 1.096 | 0.273 |
| | Synchronization - neutral | -1.2701 | 0.2975 | -4.269 | <0.001 |
| | reverse:neutral | 1.4276 | 0.3840 | 3.717 | <0.001 |
| IP2 | (Intercept) | 0.3460 | 0.2075 | 1.668 | 0.095 |
| | Order - reverse | 1.1403 | 0.2753 | 4.142 | <0.001 |
| | Synchronization - neutral | -0.0510 | 0.2413 | -0.211 | 0.833 |
| | reverse:neutral | 0.5857 | 0.4103 | 1.428 | 0.153 |

$$Model : NP2referent \sim OrderOfMention * Synchronization + (1|subject)$$
$$+(1|item), family = binomial(link = "logit")$$

24.90; $p < 0.001$ and Synchronization: $\chi^2(1) = 5.83; p < 0.05$). Participants generally inspected the NP2 referent more frequently when gaze was synchronized than when it was neutral. Moreover, model reduction revealed a significant interaction of the two predictors Order of Mention and Synchronization ($\chi^2(1) = 14.08; p < 0.001$). That is, the effect of Order of Mention varied depending on the Synchronization: Firstly, the neutral gaze condition reveals that Order of Mention by itself affected people's visual attention. In the reverse-neutral condition, the NP2 referent was inspected significantly more often than in original-neutral. We argue that this effect is due to the NP2 referent being central and being additionally highlighted as the robot initially looked downwards. Secondly, the graph also reveals that the peripherally located object (NP2 referent in original order) was inspected more often when gaze was synchronized (original-synchronized) than when it was neutral (original-neutral), suggesting that a gaze cue in original (coherent) order helped people to visually anticipate the NP2 referent. In contrast, gaze cues in reverse order did not affect the inspections on the NP2 referent (central object) compared to reverse-neutral. Instead, the NP2 referent was rather frequently inspected in reverse order even when robot gaze was neutral (during the utterance). This indicates that the central object was indeed more salient than the peripheral object.

In IP2, Order of Mention had a main effect on inspection probabilities ($\chi^2(1) = 51.99; p < 0.001$). That is, during mentioning of NP2 noun, people inspected the NP2 referent more frequently in reverse order than in original order which was also the case in Experiment 4. As before, we suggest that this reflects people visually attending more closely to the mentioned object when the referring expression required more effort to be resolved.

### Response Time

Model reduction revealed a significant interaction of both predictors, Synchronization and Order of Mention ($\chi^2(1) = 16.85; p < 0.001$). Consequently, we included both in the model fitted to our response time data. The details of this model are provided in Table 5.6. Even though Synchronization and Order of Mention had a marginal main effect on the data, the interaction of both factors was clearly more relevant for interpretation. Before interpreting the interaction, we provide pairwise comparisons here which reveal the following significant differences: Between reverse-neutral and reverse-synchronized ($p < 0.001$), reverse-synchronized and original-synchronized ($p < 0.05$), reverse-neutral and original-neutral ($p < 0.001$) and a marginally significant difference between original-synchronized and original-neutral ($p = 0.07$).

As already indicated by the inspection data in IP1, order of references in a sentence affected responsive behavior. This was further reflected in response times in both neutral conditions: People were significantly faster to validate the robot's utterance in the reverse-neutral condition than in original-neutral. This result is consistent with the findings on visual anticipation of the NP2 referent (for neutral gaze), i.e., when order was reversed people anticipated the NP2 referent, when order was original they hardly did. This suggests that reverse order of mention sentences were generally *easier* to process than original order of mention. Crucially, however, synchronization of gaze cues reversed this effect: Participants were significantly slower when gaze was synchronized and in reverse order (resulting in concurrent but conflicting referential cues) than when gaze was synchronized and in original order (concurrent and coherent order of cues).

The neutral condition was intended as a baseline for evaluating the effect of gaze synchronization while accounting for possible variations due to the manipulated Order of Mention. Interestingly, instead of observing similar behavior in each neutral

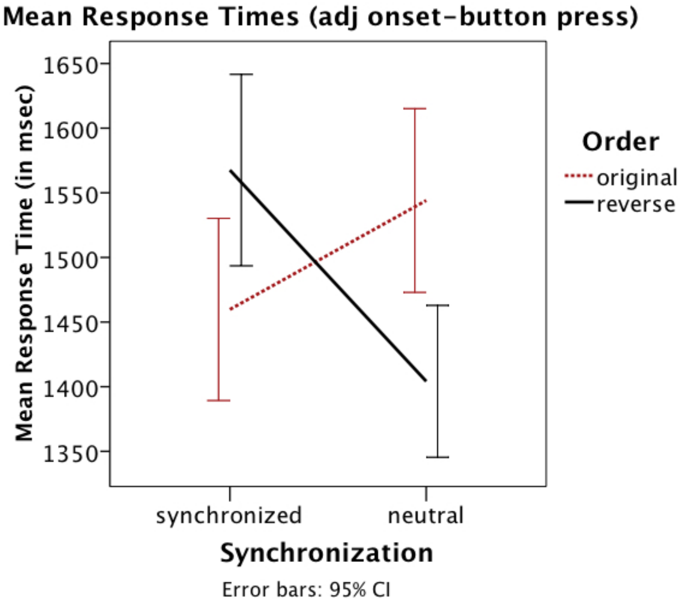


Figure 5.11.: Average response times as a result of two manipulations, Order of Mention and Gaze Sychronization.

condition, we found that reverse sentence order was *easier* to process. Despite this advantage of reverse order of mention, the synchronization of (reverse) robot gaze cues disrupted people, whereas synchronized (coherent) gaze cues in original order of mention significantly enhanced response time of this sentence order. The results for synchronized robot gaze may therefore be interpreted with respect to gaze and speech cue synchronization only: Synchronizing (reversed) gaze cue with reverse order of mention *increased* response times, while synchronizing (coherent) gaze cues with original order of mention *reduced* response times, when each is compared to its neutral gaze baseline.

Notably, even though we did not expect this effect of Order of Mention, it also supports the interpretation of response time results in Experiment 4. Previous results left open whether the main effect of Order of Mention was elicited by the order of mentioned references in the sentence or rather the chronological match of the visual and linguistic (referential) cues. The findings above suggest that reverse order facilitated comprehension and was therefore *not* the cause for increased response time observed in reverse conditions in Experiment 4. In contrast, the presented findings not only support the claim that reversed referential cues disrupt comprehension, but suggest that

Table 5.6.: Model fitted to response time data.

| Predictor | Coefficient | *SE* | t-value |  |
|---|---|---|---|---|
| (Intercept) | 1475.79 | 55.19 | 26.741 |  |
| Order - reverse | 96.24 | 40.36 | 2.384 |  |
| Synchronization - neutral | 68.89 | 39.60 | 1.740 |  |
| reverse:neutral | -230.67 | 55.94 | -4.124 |  |

|  | Coefficient | MCMCmean | pMCMC | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | 1475.79 | 1474.78 | 0.0001 | <0.001 |
| Order - reverse | 96.24 | 96.99 | 0.0156 | <0.05 |
| Synchronization - neutral | 68.89 | 69.88 | 0.0752 | 0.083 |
| reverse:neutral | -230.67 | -232.09 | 0.0001 | <0.001 |

$Model : RT \sim OrderOfMention * Synchronization + (1|subject) + (1|item)$

this disruption may indeed be greater in magnitude than our studies reveal.

Surprisingly, people's response time in condition reverse-synchronized did not directly reflect what their eye movement behavior suggested. Even though participants were slowest in this condition and fastest in condition reverse-neutral to validate the robot utterance, eye movements were extremely similar in these conditions. The eye movements plotted in Figures 5.8 and 5.9 show that people similarly anticipated NP2 in both conditions, reverse-neutral and reverse-synchronized (although gaze prior to NP2 referred to NP1 referent). Despite this apparent *visual* anticipation, the resolution of the sequential conflict seemed to induce a higher cognitive load such that people looked at the correct object but needed more time to fully grasp the meaning of all available cues. This result shows that overt visual attention does not necessarily reveal which processes caused a visual attention shift, nor whether the fixated object is actually *anticipated* as a referent for the next referring expression thereby facilitating reference resolution.

It may further be possible that user adaptation to the presented stimuli influenced the observed response time effects. Since videos in this experiment always showed

Table 5.7.: Mean response times in milliseconds for each experimental block, the development from block 1 to block 2, and overall.

| Order | Synchronization | Cue Combination | Block 1 | Block 2 | Block1-2 | Overall |
|---|---|---|---|---|---|---|
| original | synchronized | (coherent) | 1470 | 1446 | 24 | 1459 |
| original | neutral | – | 1576 | 1518 | 58 | 1544 |
| reverse | synchronized | (incoherent) | 1656 | 1485 | 171 | 1567 |
| reverse | neutral | – | 1479 | 1349 | 130 | 1404 |

synchronized, reversed or neutral robot gaze, people may have learned that robot gaze predicts referents even though it is not correctly synchronized with the utterance word order. Thus, potentially even larger effects may have been covered by participants' learning performance. To assess the influence of adaptation on response times, we conducted an additional analysis across and within each of two experimental trial blocks. The response time means for each block as well as the overall means can be found in Table 5.7. Model reduction as well as fitting the final model with three predictors (Order of Mention, Synchronization, Block) revealed a main effect of Block (Coeff. $=-79.81$, $SE$ = 28.57 , t-value $=-2.79$). That is, people were significantly faster to respond in the second block which suggests that people have generally adapted to the stimuli. The absence of any interaction between this predictor and the remaining two predictors, however, indicates that the general pattern did not change significantly over the course of the experiment. Nevertheless, the mean response times reflect an adjustment of participants to the reverse order trials in particular. While in block 1, for instance, original-synchronized was significantly faster than reverse-synchronized ($p < 0.05$), this difference disappears in block 2.

Both eye movements as well as the response time results suggest that people found it more difficult to resolve referring expressions – and ultimately comprehend the robot's utterance – when the sequential order of gaze cues was inconsistent with the order of referring expressions. Since the temporal synchronization of gaze with sentences in original order resulted in coherent (i.e., congruent) gaze and speech behavior, response times decreased – similar to the effect of congruent gaze and speech in Experiment 2. In contrast, synchronizing gaze with sentences in reverse order resulted in a reversed sequence of referential cues and, thus, response times increased. This suggests that robot gaze, even though relevant and to mentioned objects, disrupts people when it is reversely synchronized. People seem to neither ignore this kind of gaze (which would have resulted in equal response times for reverse-neutral and reverse-synchronized conditions), nor are they able to establish a mapping of the final gaze cue and the already mentioned/gazed at objects in order to predict the NP2-referent (which would have resulted in a facilitating effect of reversed gaze).[1] The reversely synchronized gaze cues rather elicit expectations about future referents which conflict with the actual utterance. Even though the referential information provided by these gaze cues is somewhat relevant, the resolution of these conflicts between expectations and the utterance is obviously demanding and slows people instead of facilitating utterance comprehension.

### Summary

To summarize the results from Experiments 4 and 5, we found that substantial temporal shifts of robot gaze with respect to its 'natural' synchronization do not affect the utility of the gaze cues whereas the linear order of the cues does. This contradicts the predictions derived from the Visual Account. The Intentional Account, in contrast, provides a plausible explanation for these results: The precise temporal synchronization is not critical since people interpret and use robot gaze as a cue to the robot's intentions rather than as a purely visual cue. The order of cues, however, affects the the utility of gaze since the order of inspections reflects the speaker's intentions regarding order of mention (Griffin and Bock, 2000; Griffin, 2001). Thus, people seem to expect that the inferred referential intentions be realized in the corresponding order. If this expectation is not met, gaze cues even disrupt comprehension, as the comparison with neutral gaze suggests. Consequently, the presented evidence for the flexible use of robot gaze cues during utterance comprehension further supports the hypothesis that people assign attentional and intentional states to robot gaze. That is, people seem to indeed establish *joint attention* with the robot and apply at least response level 3 as introduced in Section 1.4).

[1] Even though block analysis provides some evidence that people do learn to use reverse gaze.

# 6. General Discussion

We begin this chapter with a review of our result in Section 6.1 and continue with the implications and contributions these results offer to the development and improvement of human-robot interaction in Section 6.2. Subsequently, we discuss implications of the presented studies more generally for cognitive accounts of gaze processing as well as joint and shared attention. We continue with an outlook on future work in Section 6.3 before concluding this thesis in Section 6.4.

## 6.1. Summary of Results

Experiments 1 and 2 revealed that participants follow robot gaze when it is available and that they use it to resolve referring expressions. This behavior was observed even though the task neither required participants to pay attention to robot gaze, nor did gaze cues statistically help participants to predict mentioned referents across the course of an experiment (gaze effectively predicted a referent in only 55.5% of all trials). In Experiment 1, people were confronted with referentially ambiguous statements that were accompanied by referential robot gaze. People's eye movements as well as response time results suggested that people followed robot gaze to the scene referents, even when there was a visual competitor compatible with the referring expression. In Experiment 2, we manipulated the congruency between the spoken reference and the referential cues provided by robot gaze. That is, the robot either looked at an object it was about to mention (congruent), it looked an one object but then mentioned another one (incongruent), or it showed neutral gaze. We found that robot gaze that was congruent with the uttered sentence helped human interlocutors to faster validate utterances compared to when robot gaze was neutral. On the other hand, when robot gaze was incongruent with the utterance, i.e., it identified an object other than the mentioned one, people were even slower than in the neutral condition.

We hypothesized that this facilitation/disruption effect of referential robot gaze relative to neutral gaze was due to the assignment of intentional states to the robot (Intentional Account). That is, we suggested that people inferred referential intentions from robot gaze, just as has been shown to be the case for human gaze (Baron-Cohen et al., 1995; Hanna and Brennan, 2007), such that they expected an object fixated by the robot to be mentioned next. Incongruent gaze, thus, entailed a revision of expectations. An alternative explanation of the observed effects concerned the fact that gaze cues may reflexively draw an observer's visual attention to the cued direction (Visual Account). Thus, robot gaze towards the subsequently mentioned object may have facilitated reference resolution simply because listeners' attention was already on the relevant object. Incongruent gaze, in contrast, drew people's attention to one object while the spoken reference drew attention to another object. The additional shift of visual attention may have prolonged the time needed to fixate the referent, comprehend the whole utterance and respond.

In order to help decided between the Visual and the Intentional Account, we changed the task from a response time task in Experiments 1 and 2 to a production task in Experiment 3. That is, we showed participants the same stimuli as in Experiment 2 and asked them to verbally correct false robot utterances. Without time pressure on people's responses, a shift of visual attention itself to a (ir)relevant object could not explain any congruency effects such that the Visual Account could most likely be rejected. Instead, congruency effects could be explained by the Intentional Account, i.e., that robot gaze influences which referent people thought was 'intended' by the robot. And indeed, the correction statements participants produced confirmed that their correction (and which object they decided to mention) was influenced by robot gaze.

Results from Experiments 1-3, thus, provided evidence in favor of the Intentional Account. Importantly, in these first three experiments, we presupposed that robot gaze would only be helpful if it was aligned to the utterance in a human-like manner. However, our manipulations allowed no claims about the relevance of alignment of gaze and speech for the observed effects. We hypothesized that the importance of alignment may essentially depend on whether robot gaze reflects intentional states or not. Specifically, we argued that, under the Visual Account, human-like alignment robot gaze and speech would be necessary for gaze to be beneficial since this way listeners' attention would be drawn to the right object at the right moment. A temporal shift of gaze relative to speech would therefore reduce the benefit of gaze. In contrast, the Intentional Account would allow for a more flexible use of gaze since referential intentions are relevant to the sentence as a whole, and likely more persistent.

In Experiments 4 and 5, we investigated the flexibility with which people interpret and use gaze by manipulating the temporal synchronization of gaze and speech as well as the relative ordering of referential gaze cues and referring expressions. Experiment 4 revealed that the precise temporal synchronization was not critical for the utility of robot gaze. That is, a substantial temporal shift of the gaze cues such that they completely preceded the utterance (4.3 earlier than 'natural' occurrence) did not affect the general (facilitating or disruptive) influence of gaze. Interestingly, while the precise temporal synchronization was not crucial, the relative ordering of gaze and speech cues did affect response times, i.e., a reversed sequence of referential cues led to increased response times. This suggests that the inferred (referential) intentions were maintained over several seconds but that listeners were sensitive to the order of their occurrence and their realization. This is not surprising since the order of gaze cues is known to reflect the speaker's intentions regarding order of mention (Griffin and Bock, 2000; Griffin, 2001). Thus, people seem to expect that the inferred referential intentions be realized in the corresponding order during the utterance. If this expectation is not met, gaze cues – despite identifying intended referents – disrupt comprehension.

Together these results suggest a "utility spectrum" of speaker gaze as depicted in Figure 6.1: Most useful is congruent gaze that is closely aligned to speech. Gaze cues that appear in the same order as referring expressions appear similarly useful even when completely preceding the utterance. However, as evidence for this similarity is provided through the absence of a response time effect, further investigation is required to confirm this initial result.

Moreover, it was found that both incongruent gaze as well as gaze cues in reverse order to the mentioned references disrupt comprehension and slow response times compared to neutral gaze. While incongruent gaze identified an object that was not mentioned at all, reversed gaze cues identified objects that *were* mentioned in the corresponding utterance but not in the expected order.

These observations suggest that referential gaze invariably influences utterance comprehension where people infer referential intentions from robot gaze and its order of occurrence which elicit expectations about the robot utterance. That is, people attribute attentional and intentional states to the robot such that they follow its gaze to establish joint attention, typically a very useful and natural process which people clearly do not disengage from, even when gaze often has a disruptive effect. A (mis)match between the expectations constructed when jointly attending to an object and the uttered referring expressions, thus, determine whether robot gaze facilitates or disrupts comprehension.

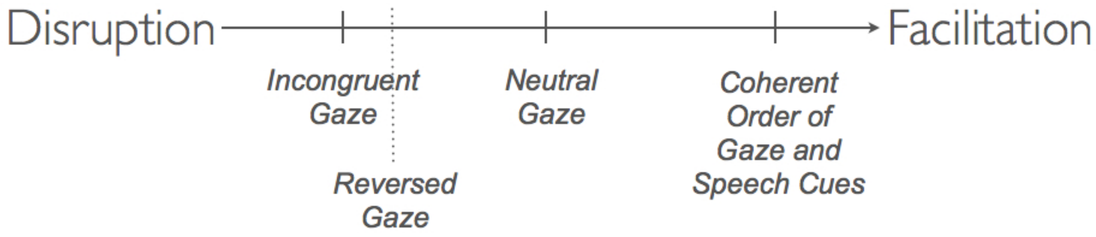


Figure 6.1.: A utility spectrum for the production of various gaze patterns shows that gaze always affects utterance comprehension – for the better or worse.

The finding that people assign intentional states to robot gaze supports the hypothesis that robot gaze is indeed processed and interpreted in a similar manner as human gaze. That is, the presented evidence provides reason to believe that further insights on how people integrate referential robot gaze during reference resolution contributes to our more general understanding of the role of gaze for grounding and disambiguating utterances in face-to-face interaction.

## 6.2. Results in Context

### 6.2.1. Robots and Virtual Characters

One might argue that the presented findings have only limited validity since we employed a remote (tele-present) robot and only allowed for minimal interaction. We argue, however, that if such behavior is elicited in such an artificial situation, it is even more likely to occur in more natural and interactive settings. Employing a virtual agent rather than a robot is one possible alternative but it is worth noting that there are differences between these two types of agents that could potentially affect the usage of gaze. While a virtual agent "lives" in its own world and may be assumed to have complete knowledge of its environment, a robot shares the environment with a human interlocutor and is not expected to have full knowledge of the world. This may lead to different expectations and impressions of an agent versus a robot. Mistakes and errors are potentially more acceptable and less irritating when communicating with a robot. Moreover, robots seem to elicit more anthropomorphic interaction and attributions than agents Kiesler et al. (2008). Kiesler and colleagues investigated, for instance, whether a robot and a virtual agent in co-present and remote conditions are perceived differently. The results of their questionnaire, which participants had to fill out after a 10-15min discussion with the robot/agent, indicate that the robot was perceived as more life-like, having more positive personality traits and being liked better. However, whether the robot/agent was co-present or remote, i.e. recorded and projected onto a screen, did not seem to substantially affect participants' impressions.

The reported findings are of considerable importance for the design of systems controlling robot gaze. We have shown that referential robot gaze contributes to more rapid understanding and, thus, is to be preferred over a robot that does not look at the objects it is talking about. However, when the location of a referent (or which object to look at) cannot be determined (incongruent gaze may be the consequence) it is advisable not to initiate robot gaze movements since these may disrupt the comprehension of the user. Similarly, the order of references in the utterance should be considered when initiating gaze movements to referents in the scene in order to avoid disruption by incoherent robot gaze. Moreover, our results suggest that a precise temporal synchronization is otherwise not essential for robot gaze to be beneficial. This is especially important since precise temporal alignment is difficult to achieve as it depends on issues such as knowing which nouns could and should be accompanied by referential gaze or considering both the velocity of head or eye movements and the required duration.

We further suggest that the presented experimental setting is suited to the more general investigation of beliefs humans have about robots and their capabilities. The attribution of beliefs, goals and desires to others is a crucial skill in social interaction (Baron-Cohen et al., 1997a, 1985). This capability is necessary in order to realize, 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 be as competent as a person, but with increased communicational skills the expectations towards the robot will also rise.

### 6.2.2. A Theory of Mind with Robots

With the increased competence and improved appearance of robots and other artificial agents, it is not surprising that researchers have begun to consider the utility of theory of mind (ToM) models for human-robot interaction. One approach concerned the improvement of the robot's competence to communicate with people. Scassellati (2000), for instance, attempted to implement two ToM models on a robot system. He outlined the long-term research goal 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. Another approach concerned the investigation of what mental models people have for robots with a focus on the appearance of the robot and the anthropomorphization of it. Kiesler and Goetz (2002), for instance, examined and compared the impressions people obtained from a robot that featured a visible hardware component including cables versus a robot whose hardware was hidden. A questionnaire revealed that the visible hardware caused participants to consider the robot less "reliable" and more "powerful". In another study, Hegel et al. (2008) investigated the effect of an agent's appearance and the associated stereotypical knowledge in an FMRI-study. Participants were asked to play the prisoner's dilemma with each of the following four partners: A person, a humanoid robot, a functional lego robot consisting of mechanic arms operating a keyboard, and a computer. Importantly, each partner gave completely randomized responses so that the only difference between each interaction was the belief of who participants thought they interacted with. The FMRI-data revealed increased activity during all interaction in those brain regions typically considered to be participating in common ToM tasks. Results further showed a tendency towards higher activity in the respective brain region when participants were facing robotic partners relative to the laptop partner. In a different study, Groom et al. (2009) also showed that people seem to attribute identity to robots with a humanoid form, crucially an identity that is separate from that of their own. In contrast, people perceived a robot with a car form potentially as team member or even as an extension of themselves. Thus, appearance and anthropomorphism clearly affect the expectations people have of their partners. However, such studies do not capture the expectations and a potential ToM (and the adaptation thereof) based on the partner's actual *behavior* during an interaction.

We argue that our approach involving behavioral measures in a human-robot interaction scenario allows us to investigate in detail whether humans build a ToM for a robot they interact with and what the nature of this ToM is. People's behavior can be linked precisely to the robot's behavior, thus, potentially providing insights into an incremental construction and enrichment of a ToM based on behavior rather than general appearance (which was in fact not manipulated at all).

Recall that having a theory of mind means to possess and use knowledge about mental states in general, and about epistemic mental states (believing, knowing, pretending) in particular, and to use this knowledge in a "theory-like" way to reason about and predict actions of others (Baron-Cohen, 1995, p.51ff). That is, a person can draw inferences about why another person, or robot in this case, behaves the way she does because she can imagine what goals and intentions have elicited this behavior. People's use of robot gaze as a cue to the robot's referential intentions reveals their understanding of the robot as an intentional being with perceptual and volitional states. Further studies could potentially provide answers to more specific questions such as: What do people think about the robot's cognitive capabilities? Which modality do people preferably trust in and consider more reliable? Knowing what ToM model people construct for a robot benefits robot development in general and user adaptive behavior in particular.

### 6.2.3. The Role of Appearance (versus Behavior)

The robot we used in our studies had a very simple appearance with almost no anthropomorphic features. A stereo-camera mounted on a pan-tilt unit served as head and eyes simultaneously. This camera was the only moving part of our robot and only through this movement the robot appears as actively performing. Yet, we observed participant behavior that is very similar to what Hanna and Brennan (2007) observed. In their studies, listeners rarely looked at the speakers' face to detect where the speaker was gazing at and rather used the speaker's head orientation peripherally. We interpret this is as additional support for the claim that robot gaze as a combination of head orientation and gaze can in principle be used similarly to human speaker's gaze, even though the robot has no anthropomorphic eyes. This is not so surprising as Emery (2000) suggests that the eyes are only the first choice for interpreting an individual's direction of attention but not the only one. Instead he describes a hierarchy of cues (gaze, head, body) the use of which depends on their availability.

Related evidence for the importance of the camera movement (rather than the appearance) is provided by studies that explicitly investigated the role of motion for the assignment of goals and intentions to moving entities (e.g. Heider and Simmel, 1944). Using a simple animation which showed moving geometrical figures, Heider and Simmel (1944) found that those movements were often interpreted as one object hitting the other, as pushing or pulling actions or as leading and chasing events. That is, people interpreted movements of simple geometric shapes as goal-driven events, with one entity as agent and another as patient, and even hypothesized about *motives* for these events, suggesting that they ascribed goal and intentions to the moving objects.

This suggests that people do not only rely on the anthropomorphic appearance of a face and/or eyes to elicit 'natural' reactions towards an entity, but that certain movement patterns, potentially only with appropriate scope and timing, can achieve this as well. Whether the appearance is closer to a pair of eyes or a complete head may not play a particularly significant role here. In our case it seems that the camera *movement* which is aligned to the robot's utterance is in fact the reason why people attribute cognitive functions to it. While reflexive attention towards the robot camera may explain immediate gaze-following (see following section), we propose that it is the attribution of cognitive functions (based on plausible motion) which ultimately explains why we observe an effect of robot gaze on reference resolution/intention recognition.

For these reasons, we feel there is considerable reason to believe that our results allow conclusions about general mechanisms involved in gaze processing. Moreover, our experimental setting offers several methodological advantages. Hanna and Brennan's studies, for instance, have focused on people's flexibility in interpreting gaze direction by forcing listeners to re-map speakers' gaze to their own (different) object arrangement. While their aim was to investigate whether and how flexible a gaze cue is, our studies focus on examining the integration process of referential information provided by gaze and speech, especially in cases of mismatch. By using a robot as interaction partner, we can create plausible mismatching references by introducing wrong or erroneous (i.e., incongruent) robot behavior. In such cases, re-mapping of perceived gaze is not appropriate and cannot help to combine cues to one consistent reference. Instead, people have to make sense of the information they perceived by actively weighing one cue (or one modality, i.e., speech versus gaze/vision) higher than the other and eventually make a decision based on that. Such a design can therefore provide insights about the contribution of gaze relative to speech. Thus, benefits of our HRI design are, on the one hand, that the robot (or agent in general) can be used to produce behavior that is almost arbitrarily variable while, on the other hand, experiments can be controlled for various factors such as balancing of stimuli, which errors should occur and when, as well as cue validity. That is, we can induce errors where necessary and eliminate undesired behaviors that may mask other effects. For this reason, we believe that the behavior observed in human-robot interaction may indeed yield insights that are as valuable to cognitive research as they are to the development of human-machine interfaces.

### 6.2.4. Endogenous and Exogenous Attention Shifts

Previous research has increasingly involved computerized paradigms to explore the role of gaze and other cues to control attention. A number of such studies suggest that schematized gaze cues (schematic face, stylized pair of eyes, photographs of a face etc.) attract people's attention to the cued location in an automatic manner (Friesen and Kingstone, 1998; Driver et al., 1999; Langton and Bruce, 1999). That is, gaze - and to some extent also other direction-giving cues such as arrows (Bayliss and Tipper, 2005; Tipples, 2008; Marotta et al., 2009) - are assumed to trigger a reflexive (exogenous) attention shift, rather than having people voluntarily orient towards the indicated direction (endogenous attention shift). Friesen and Kingstone (1998) found, for instance, that participants were faster to detect, localize and identify (within-subject task variation) a target stimulus when the eyes of a centrally presented face cued the target's location. This cueing effect was found only for a relatively short time window, i.e., when the target stimulus appeared 105, 300 or 600ms after the cue onset, and seemed to disappear when target stimulus appeared only 1,005ms after the cue onset. Moreover, it was found that reaction times were equal among the "uncued" and "neutral" conditions, i.e. when the eyes cued the opposite location and when they were directed straight ahead. The authors therefore suggest that "*gaze direction is producing an attentional benefit (RT at the cued location < RT at the neutral location) with no attentional cost (RT at the neutral location = RT at the uncued location).*" (Friesen and Kingstone, 1998, p.493f).

The reported results may hold for direction-giving cues that only elicit reflexive attention shifts but this conclusion seems to underestimate the influence of on-line gaze that reflects attentional or even intentional states. Despite the evidence suggesting that these gaze cues simultaneously elicit voluntary attention shifts to a limited extent (Tipples, 2008), it is questionable whether this is comparable to the level of volition involved in joint attention, for instance. On the basis of previous research on joint attention and our results presented in this thesis, we argue that on-line gaze behavior, in contrast to static gaze cues, expresses attentional focus and communicative intentions after all (Intentional Account). That is, robot gaze not only happens to attract people's visual attention to a target which is then mentioned, but it also conveys information about what the robot presumably intends to mention next. We suggest that this is the reason why both incongruent (similar to Friesen's "uncued" condition in Friesen and Kingstone, 1998) and reverse robot gaze disrupt utterance comprehension, in contrast to the static gaze cue used by Friesen and Kingstone (1998).

The hypothesis that gaze orients attention through endogenous processes, is supported by evidence reported in a study of patient 'EVR' who suffers from neural impairments in the frontal lobe (Vecera and Rizzo, 2006). In this study, EVR was presented with a number of different cues (peripheral, gaze, word) each potentially cueing a target stimulus. The patient was shown to be able to detect peripheral cues that further facilitated target detection, suggesting intact exogenous attentional processes. However, for centrally presented cues such as words ("left","right") or eye gaze (schematic face) EVR showed no cueing effect which suggests that these cues do not trigger attention reflexively. Instead, gaze and words seem to share voluntary (endogenous) attentional processes that are disrupted in EVR.

Thus, previous results suggest that gaze cues do not (only) elicit exogenous attention shifts but direct visual attention in a voluntary (endogenous) manner. It remains unclear, however, to what extent voluntary attention shift can be related to the intentions people assign to (robot) speaker gaze such that they construct expectations about utterance content.

Additional support for voluntary attention shift comes from a number of studies on intentional gaze processing (Castiello, 2003; Bayliss et al., 2006; Becchio et al., 2008). Bayliss et al. (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". Our data support this view that gaze is indeed processed as an intentional cue, suggested by Becchio et al. (2008). Moreover, our results suggest that intentional gaze processing is applied not only to human eyes but also when faced with an extremely simple realization of robot gaze (represented by a moving camera).

### 6.2.5. Gaze and Situated Language Comprehension

In this section, we briefly outline where we see contributions of our results to existing theories and findings on how people comprehend language in a certain visual context and ground utterances in the environment.

### The Coordinated Interplay Account

The Coordinated Interplay Account (CIA) proposed by Knoeferle and Crocker (2006) was designed to explain human gaze behavior and the influence of attended scene information during situated comprehension. The CIA builds on findings from the visual world paradigm and states that incremental interpretation of utterances directs visual attention towards mentioned and anticipated objects and events in the scene. The obtained visual information, in turn, further influences interpretation of these utterances. That is, the CIA consists of three informationally and temporally dependent stages: incremental sentence comprehension, utterance-mediated visual attention shifts, and the integration of attended scene aspects and current sentence interpretation.

Our findings on eye movements are broadly consistent with Knoeferle and Crocker's CIA, but require an extension to that model such that it accommodates the multimodality of the utterance itself which consists of a spoken message (the sentence) as well as – even if unintended – a visual component (speaker gaze). As we have shown, both utterance and speaker gaze direct people's visual attention in the scene and are used to ground utterance meaning during comprehension. That is, while speaker gaze could be considered a part of the spoken message it accompanies, it is also part of the visual scene and, thus, information which the listener obtains visually during comprehension and links to utterance interpretation. Specifically, the gaze component of such a multi-modal message serves to ground utterance meaning in the visual scene.

Interestingly, just as Knoeferle and Crocker (2006; 2007) show that scene events can override 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 enforces the requirement that an appropriate model acknowledges and explains the multi-modality of both the message and also the receiver. The receiver, or listener, also perceives and processes information obtained through different channels which she then needs to integrate into a coherent message.

### Interpretation of Speaker Gaze Varies

When comparing Hanna and Brennan's (2007) results with our findings, two interesting issues arise that are worth discussing here. Firstly, Hanna and Brennan noted that the presentation order of blocks (with one experimental condition each) had an effect. That is, if block with congruent trials was first and incongruent trials came second, people were better and using the speaker gaze for early reference resolution. This suggests that people *learned to use* speaker gaze. In contrast, if the block with incongruent trials (speaker gaze was uninformative) came first, people performed worse with respect to quick reference resolution. The authors suggested that these participants had *learned to ignore* or avoid speaker gaze since it was not useful. In our studies, however, we did not observe avoidance of robot gaze even when it was frequently misleading. Instead of showing blocks of each condition, we interleaved conditions in stimulus presentation. This possibly affected how people used gaze since utility of gaze could not be predicted. It could also be argued that the continuous use of robot gaze is related to people's patience towards a robot being larger than towards other people. Since the interaction is minimal, however, and robot gaze is simply not particularly helpful, this explanation is only partly convincing. Either way, our results suggest that gaze-following is somewhat automatic and that gaze cues are always used, otherwise we would not have observed disruption effects.

Secondly, Hanna and Brennan report gaze-following only to the extent that matchers/listeners began looking at the target side of the display 500ms after the director/speaker looked at the target (visual point of disambiguation, VPoD). Matchers only began fixating the target (and competitor) when the prenominal adjective was mentioned (over 1,500-2,000ms *after* speaker gaze to target). In contrast, we observed listeners' gaze-following that is temporally closely aligned to the speaker gaze: Inspections on an object gazed at by the robot rose immediately after speaker gaze, visible in the subsequent time window of 1,000ms. Participants seemed to clearly identify the visual referent which may be explained by the explicit and obvious orientation of the robot head/eyes.

### Speaker Gaze Influences Language Learning

Recall that speaker gaze is a useful cue the speaker's referential intentions, her desires and goals (Baron-Cohen et al., 1995, 1997b) and that children learn to interpret and use this cue at a very early stage during development (D'Entremont et al., 1997; Flom et al., 2007). It is therefore reasonable to assume that children (or adult learners) use speaker gaze to help them resolve, ground and, thus, learn unknown words. That is, gaze cues may help to acquire the meaning of words or, more generally, what a sentence is about. In particular, gaze cues could be extremely helpful in a complex language learning scenario (for both children and adults) when the word for an object, an action or event is unknown.

Accordingly, Nappa et al. (2009) conducted a study with children investigating the effect of gaze cues for unknown verb interpretation. Listeners viewed scenes depicting an action that require a perspective, i.e. one interpretation of the action selects character A as agent and character B as patient, while the complementary verb reverses this role assignment (e.g. *chase* and *flee*). Given the general preference for so-called "source-to-goal" or to-path verbs over from-path verbs, the acquisition of to-path verbs is assumed to be easy while the overlapping context in which both verbs appear thereby reduces the probability of correctly learning the from-path verb. To examine whether gaze can influence perspective selection and, thus, which verb is learned, Nappa and colleagues manipulated which character was gazed at while the speaker uttered a sentence containing an unknown verb. They found that while biased towards to-path verbs, speaker gaze indeed modulated the selection of a character as subject for the uttered sentence and, thus, the choice of the verb perspective. However, this was only the case when subject and object references where ambiguous (e.g. "He's blicking him."). In the case of unambiguous references such that subject and object were uniquely identified, children mostly took the to-path perspective to assign meaning to the unknown verb. That is, speaker gaze did not affect meaning assignment anymore even though listeners followed speaker gaze. This result seems to suggest that gaze cues do not override linguistic and conceptually preferred information, suggesting rather that people prioritize and interpret speech over gaze in case of incongruent speaker behavior.

In contrast, our results suggest that people use gaze also when utterances are unambiguous, as shown in Experiments 4 and 5. Moreover, results from Experiment 3 have shown that people sometimes consider the looked at object exclusively as intended referent and correct an initially valid robot utterance accordingly (in true-incongruent trials). The limited influence of gaze in Nappa's study (2009) could be due to the presentation mode in which the speaker/gazer is not part of the environment that contains the described event. That is, the interpretation of gaze as visual reference reflecting attentional and intentional states is less likely. However, we can only speculate about reasons and it could just as well be the case that speaker gaze has simply more importance for reference resolution (linking referring nouns to objects) than on the comprehension and grounding of verbs.

## 6.3. Future Work

In this section we first address new questions that arose from experiments and their results and suggest how these might be tackled. Secondly, we outline some more speculative ideas that have evolved during the work on this thesis and highlight connections between our research and other areas of inquiry.

### 6.3.1. Next Steps

Firstly, we would like to replicate some of the presented studies with minor changes to the manipulations. Some factors could not be optimally manipulated since we essentially employed the same set of visual stimuli across the experiments. This constrained the variation of gaze cues in terms of their direction, for instance, as well as the location of referents in the scene.

Secondly, possible explanations for inspection and response time patterns found in conditions reverse-neutral and reverse-synchronized in Experiment 5 are worth exploring in further detail. Interestingly, the inspection patterns in both conditions were extremely similar while response times suggest a fundamental difference in the effort needed to resolve the references and validate the sentences. This suggests that the eye movements reflect different underlying processes involved in the integration of linguistic and visual cues. Thus, looking at an object may not necessarily mean that the listener anticipates this object, it could also mean that the listener needs more information and/or more time to resolve a multi-modal reference and link the linguistic content with the visual reference and the corresponding object. Another issue related to eye movements and underlying processes concerns the role of covert attention. In Experiment 2, for instance, we have looked at the effect of gaze-following (i.e., overt visual attention) on response time, showing that the manipulation of robot gaze congruency had a greater effect on utterance comprehension when people actually followed robot gaze. However, congruency affected response times equally even when people did not follow robot gaze overtly. This suggests that gaze cue may be exploited covertly and integrated during reference resolution such people's eye movements do not necessarily reflect the influence of speaker gaze.

A third issue concerns the movement of the robot's head/eyes which was rather slow in the presented studies. One might argue that people consider such slow movement rather as a search action than as a gaze cue and that people track the robot's head instead of interpreting it as providing referential cues. However, we observed that people rarely looked at the robot head while remarkably accurately identifying and inspecting visual referents (looked at by the robot). Even though a faster, saccade-like gaze cue would be interesting to test and probably easier to experiment with as timing would be more precise, we have reason to believe that the gaze cues used in the presented experiments have in fact been interpreted as cues to the robot's attentional states.

Finally, we propose a follow-up study to Experiment 3 which further explores the predictions made by the Visual Account. Specifically, we envisage trials with neutral robot gaze in which a very subtle visual cue such as a flicker (as in the "Flickering Cake Paradigm" proposed by Christoph Scheepers) draws people's visual attention towards the same object at the same time at which previously robot gaze drew attention to that object. Such a manipulation would provide a means to compare between a purely visual cue, that *happens* to direct people's attention to an object before it is mentioned, with a cue which is suggested to reflect referential intentions. A difference in how these cues affect people's correction statements – and therefore which object was believed to be intended – would provide strong evidence in favor of the Intentional Account.

### 6.3.2. Ideas for Future Work

One of our future goals is to compare the nature of an interaction with a robot in contrast to the interaction with a virtual character (VC). We intend to replicate previous studies in an interaction setting with a virtual character investigating whether the same principles hold in both HCI and HRI. Moreover, it would be interesting to systematically explore to which extent participants feel that they share an environment with such a character and consider objects in the visual world versus objects in the "real world" to be common ground, for instance. Such studies could be used to examine people's beliefs about the partner's attentional and epistemic states and would provide concrete insights into the theory of mind of the person interacting with a VC. A related research question concerns the issues of what precisely referential gaze may communicate. That is, are affordances (Gibson, 1977) activated based on a gaze cue (Becchio et al., 2008) and, if so, does their activation (possibly spuriously) influence event representations? Studying gaze cues and their relation to event representation is generally conceivable in both HRI and HCI scenarios but may again reveal differences in what people hypothesize about the gazer (ToM) and what meaning they assign to the gaze cue.

Replication of our previous experiments with a VC naturally also raises questions about the role of velocity and alignment of the VC's eye movements. Since a VC can execute eye movements fast and frequently, essentially without any physical constraints, it becomes critical to decide *how fast* and *how often* the character should look at entities in order to appear as natural and remain as informative as possible. Similarly, we would expect differences in people's inspection behavior due to differences in the visibility of gaze direction: Robot gaze was so explicit and visible that people picked up this cue without having to look at the robot head. However, the gaze direction of a VC (depending also on the use of head movements) could be more difficult to pick up such that people might need to look at the character's head/eyes more frequently.

Importantly, the increased velocity of gaze in VCs opens up the possibility to include and combine several functions of gaze beyond visual references. That is, various gaze cues could be implemented and used simultaneously to fulfill additional tasks such as turn-coordination. Taking turns, however, presupposes another essential aspect of future experiments, namely greater interaction.

There are indeed cases where gaze-following and joint attention occur and are useful even in minimal interaction (of the sort created in our HRI studies), for instance, when children learn from their mother. A classical situation contains an adult that utters a description or explanation of some sort while looking at mentioned objects. Once the child has learned that eye gaze reflects visual attention and referential intentions, she can use this cue to comprehend what the adult says – even when she did not know one or more words in the utterance (Flom et al., 2007). Moreover, Nappa et al. (2009) have shown that gaze cues in such a minimal interaction scenario can help people to assign meaning to, that is, to *learn* an unknown word.

However, the application of several gaze behaviors and observing their effects on participant behavior is only possible in a scenario allowing real interaction, when speech and gaze are produced and comprehended simultaneously. In such an interaction scenario, gaze could adopt several functions: While coordinating turns, it could also, as mentioned above, be used as an additional referential cue helping language learners to ground new words, or to express "mental" states such as uncertainty or confusion about something. Crucially, in such an interaction the agent could similarly make use of the person's gaze, constraining and simplifying its own domain of interpretation when processing an utterance of the person (see e.g. Kaur, Manpreet and Tremaine, Marilyn and Huang, Ning and Wilder, Joseph and Gacovski, Zoran and Flippo, Frans and Mantravadi, Chandra Sekhar, 2003; Prasov and Chai, 2008, for some ideas on how to do this). This way, insights on people's use of visual and linguistic information to efficiently resolve references could be used to similarly facilitate robot utterance comprehension.

## 6.4. Conclusion

Summarizing the results presented in this thesis, we have provided evidence suggesting that detailed insights from situated human-human interaction (HHI) can be applied to human-robot-interaction (HRI) and that such cognitively motivated robot-gaze behavior is beneficial for HRI.

In Experiment 1, we have shown that people follow robot gaze and that referentially ambiguous utterances that are accompanied by referential robot gaze are understood equally fast as unambiguous statements. Results from Experiment 2 further suggest that people used robot gaze to anticipate an upcoming referent such that congruent robot gaze facilitated comprehension while incongruent gaze disrupted comprehension relative to neutral gaze. We hypothesized that this utility of gaze would be caused by the attentional and intentional states that people ascribe to the robot. That is, we argued that people may indeed establish joint attention with the robot, interpreting its gaze to indicate what the robot attends to and what it intends to mention. Findings from Experiment 3 confirm this hypothesis and show that robot gaze modulates which object – in the case of wrong utterances – is considered as intended referent. In Experiments 4 and 5, we have examined the relevance of alignment between gaze and speech for such an intentional interpretation of robot gaze. The results again suggest that people infer referential intentions from robot gaze such that gaze similarly affects utterance comprehension when it is correctly synchronized as when it is shifted to precede the utterance. Moreover, the order of referential cues has been found to affect comprehension, indicating that people interpret gaze cues in their occurring order and expect the retrieved referential intentions to be realized accordingly. Thus, our findings converge to the result that people establish joint attention with a robot and infer attentional and intentional states from its gaze, as suggested in the description of response level 3 from Section 1.4.

The video-based one-way interaction in our experiments can hardly be considered to allow *shared attention* (response level 4 in Section 1.4) such that participants and the robot are mutually aware of their attentional states and use this information to cooperate or just to share an experience. However, our results clearly suggest that people interpret robot gaze in a similar way they interpret human gaze which is evidence for the utility of our experimental paradigm for investigating not only the role of robot gaze but also aspects of gaze processing in general.

# Bibliography

Adams, R. B. and Kleck, R. E. (2003). Perceived Gaze Direction and the Processing of Facial Displays of Emotions. *Psychological Science*, 14:644–647.

Allopenna, P., Magnuson, J., and Tanenhaus, M. (1998). Using Pointing and Describing to Achieve Joint Focus of Attention in Dialogue. *Journal of Memory and Language*, 38:419–439.

Altmann, G. and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

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

Argyle, M. and Dean, J. (1965). Eye-Contact, Distance and Affiliation. *Sociometry*, 28(3):289–304.

Baayen, R., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. MA: MIT Press/Bradford Books.

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

Baron-Cohen, S., Campbell, R., Karmiloff-Smith, A., Grant, J., and Walker, J. (1995). Are children with autism blind to the mentalistic significance of the eyes? *British Journal of Developmental Psychology*, 13:379–398.

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

Baron-Cohen, S., Wheelwright, S., and Jolliffe, T. (1997b). Is There a "Language of the Eyes"? Evidence from Normal Adults, and Adults with Autism or Asperger Syndrom. *Visual Cognition*, 4:311–331.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, 59(4):457 – 474. Special Issue: Emerging Data Analysis.

Bates, D. (2005). Fitting linear mixed models in R. *R News*, 5:27–30.

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

Bayliss, A. and Tipper, S. (2005). Gaze and arrow cueing of attention reveals individual differences along the autism spectrum as a function of target context. *British Journal of Psychology*, 96:95–114.

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

Berry, D. S., Misovich, S. J., Kean, K. J., and Baron, R. M. (1992). Effects of Disruption of Structure and Motion on Perceptions of Social Causality. *Personality and Social Psychology Bulletin*, 18:237–244.

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

Breazeal, C., Kidd, C., Thomaz, A., Hoffman, G., and Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, pages 708–713.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, 97:523–547.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H., and Yan, H. (1999a). Embodiment in Conversational Interfaces: Rea. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'99)*, pages 520–527.

Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., editors (1999b). *Embodied Conversational Agents*. MIT Press.

Cassell, J. and Thórisson, K. (1999). The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. *Applied Artificial Intelligence*, 13:519–538.

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

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

Chambers, C. G., Magnuson, J. S., and Tanenhaus, M. K. (2004). Actions and Affordances in Syntactic Ambiguity Resolution. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30:687–696.

Clark, H. H. and Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1):62–81.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6(1):84–107.

Crawley, M. J. (2007). *The R Book*. Wiley Publishing.

Crocker, M. W., Knoeferle, P., and Mayberry, M. ( in press). Situated sentence processing: The coordinated interplay account and a neurobehavioral model. *Brain and Language*.

D'Entremont, B., Hains, S., and Muir, D. (1997). A Demonstration of Gaze Following in 3- to 6-Month-Olds. *Infant Behavior and Development*, 20:569–572.

Desimone, R. and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–222.

Dovidio, J. F. and Ellyson, S. L. (1982). Decoding Visual Dominance: Attributions of Power Based on Relative Percentages of Looking While Speaking and Looking While Listening. *Social Psychology Quarterly*, 45:106–113.

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

Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology*, pages 283–292.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., and Tanenhaus, M. K. (1995). Eye Movements as a Window into Real-Time Spoken Language Comprehension in Natural Contexts. *Journal of Psycholinguistic Research*, 24:409–436.

Emery, N. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, 24:581–604.

Flom, R., Lee, K., and Muir, D., editors (2007). *Gaze-Following: Its Development and Significance*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

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

Friesen, C., Ristic, J., and Kingstone, A. (2004). Attentional Effects of Counterpredictive Gaze and Arrow Cues. *Journal of Experimental Psychology: Human Perception and Performance*, 30(2):319–329.

Frith, C. and Frith, U. (2005). Theory of Mind. *Current Biology*, 15:R644–R646.

Gibson, J. J. (1977). The theory of affordances. In Shaw, R. and Bransford, J., editors, *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, pages 67–82. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ.

Greenbaum, P. E. (1985). Nonverbal Differences in Communication Style between American Indian and Anglo Elementary Classrooms. *American Educational Research Journal*, 22:101–115.

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

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

Groom, V., Takayama, L., Ochi, P., and Nass, C. (2009). I Am My Robot: The Impact of Robot-building and Robot Form on Operators. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction (HRI '09)*, pages 31–36. ACM.

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

Hegel, F., Krach, S., Kircher, T., Wrede, B., and Sagerer, G. (2008). Theory of mind (ToM) on robots: a functional neuroimaging study. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction (HRI '08)*, pages 335–342. ACM.

Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, 57:243–259.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11):498–504.

Hietanen, J. K., Leppänen, J. M., Peltola, M. J., Linna-aho, K., and Ruuhiala, H. J. (2008). Seeing direct and averted gaze activates the approach-avoidance motivational brain system. *Neuropsychologia*, 46:2423–2430.

Hoffman, J. E. and Subramaniam, B. (1995). The role of visual attention in saccadic eye movements . *Perception & Psychophysics*, 57:787–795.

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

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4):434 – 446. Special Issue: Emerging Data Analysis.

Kanda, T., Ishiguro, H., and Ishida, T. (2001). Psychological Analysis on Human-Robot Interaction. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA) '01*, pages 4166–4173.

Kaplan, F. and Hafner, V. V. (2006). The challenges of joint attention. *Interaction Studies*, 7:135–169.

Kaur, Manpreet and Tremaine, Marilyn and Huang, Ning and Wilder, Joseph and Gacovski, Zoran and Flippo, Frans and Mantravadi, Chandra Sekhar (2003). Where is "it"? Event Synchronization in Gaze-Speech Input Systems. In *Proceedings of the 5th international conference on Multimodal interfaces (ICMI '03)*, pages 151–158. ACM.

Kendon, A. (1967). Some Functions Of Gaze-Direction In Social Interaction. *Acta Psychologica*, 26:22–63.

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

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

Kipp, M. and Gebhard, P. (2008). Igaze: Studying reactive gaze behavior in semi-immersive human-avatar interactions. In *Proceedings of the 8th International Conference on Intelligent Virtual Agents (IVA-08)*, pages 191–199.

Kliegl, R., Masson, M. E., and Richter, E. (in press). A linear mixed model analysis of masked repetition priming. *Visual Cognition*.

Kliegl, R., Risse, S., and Laubrock, J. (2007). Preview Benefit and Parafoveal-on-Foveal Effects From Word n+2. *Journal of Experimental Psychology: Human Perception and Performance*, 33:1250–1255.

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

Knoeferle, P. and Crocker, M. W. (2007). 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.

Knoeferle, P. and Crocker, M. W. (2009). Constituent order and semantic parallelism in online comprehension: Eye-tracking evidence from German. *The Quarterly Journal of Experimental Psychology*, 62:2338–2371.

Knoeferle, P., Crocker, M. W., Pickering, M., and Scheepers, C. (2005). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition*, 95:95–127.

Kreysa, H. (2009). *Coordinating speech-related eye movements between comprehension and production*. PhD thesis, University of Edinburgh.

Kreysa, H., Pickering, M. J., Haywood, S. L., and Henderson, J. M. (2009). Gaze Projection: Availability of speakers' eye movements affects listeners' comprehension of object descriptions. Poster at the 22nd CUNY Conference on Human Sentence Processing, Davis CA.

Kylliäinen, A. and Hietanen, J. K. (2004). Attention orienting by another's gaze direction in children with autism. *Journal of Child Psychology and Psychiatry*, 45:435–444.

LaFrance, M. and Mayo, C. (1976). Racial Differences in Gaze Behavior During Conversations: Two Systematic Observational Studies. *Journal of Personality and Social Psychology*, 33:547–552.

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

Langton, S. R., Watt, R. J., and Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Science*, 4:50–59.

Leekam, S. R., Hunnisett, E., and Moore, C. (1998). Targets and Cues: Gaze-following in Children with Autism. *Journal of Child Psychology and Psychiatry*, 39:951–962.

Leekam, S. R., Lopez, B., and Moore, C. (2000). Targets and Cues: Gaze-following in Children with Autism. *Developmental Psychology*, 36:261–273.

Mannan, S., Ruddock, K., and Wooding, D. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, 10:165–188.

Marotta, A., Casagrande, M., Raffone, A., Martella, D., Sebastiani, M., and Maccari, L. (2009). Gaze and arrow induce different effects on attentional orienting as a function of target context. In Taatgen, N. and van Rijn, H., editors, *Proceedings of the 31th Annual Conference of the Cognitive Science Society*, Amsterdam, Netherlands.

Matin, E., Shao, K., and Boff, K. (1993). Saccadic overhead: information processing time with and without saccades. *Perception and Psychophysics*, 53:372–380.

Meltzoff, A. N. and Brooks, R. (2007). Eyes wide shut: The importance of eyes in infant gaze-following and understanding of other minds. In Flom, R., Lee, K., and Muir, D., editors, *Gaze-Following. Its Development and Significance*, pages 217–241. Lawrence Erlbaum Associates, Publishers, Mahwah, NJ.

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

Moll, H., Koring, C., Carpenter, M., and Tomasello, M. (2006). Infants Determine Other's Focus of Attention by Pragmatics and Exclusion. *Cognition and Development*, 7:411–430.

Moore, C. and Dunham, P., editors (1995). *Joint Attention: Its Origins and Role in Development*. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ.

Mutlu, B., Hodgins, J., and Forlizzi, J. (2006). 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.

Nappa, R., Wessel, A., McEldoon, K. L., Gleitman, L. R., and Trueswell, J. C. (2009). Use of speaker's gaze and syntax in verb learning. *Language Learning and Development*, 5(4):203–234.

Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32:3–25.

Prasov, Z. and Chai, J. Y. (2008). What's in a Gaze? The Role of Eye-Gaze in Reference Resolution in Multimodal Conversational Interfaces. In *Proceedings of the International Conference on Intelligent User Interfaces*. ACM.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences*, 4:515–526.

Ristic, J., Friesen, C. K., and Kingstone, A. (2002). Are eyes special? It depends on how you look at it. *Psychonomic Bulletin & Review*, 9:507–513.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735.

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

Schofield, T. J., Parke, R. D., Castaneda, E. K., and Coltrane, S. (2008). Patterns of gaze between parents and children in eurpoean american and mexican american families. *Journal of Nonverbal Behavior*, 32:171–186.

Schroeder, M. and 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.

Sedivy, . J. C., Tanenhaus, M. K., Chambers, C. G., and Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71(2):109–147.

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

Staudte, M. and Crocker, M. W. (2008). The utility of gaze in human-robot interaction. In *Proceedings of "Metrics for HRI", Workshop at the 3rd ACM/IEEE Conference on Human-Robot Interaction (HRI'08)*, Amsterdam, Netherlands.

Staudte, M. and Crocker, M. W. (2009a). Producing and resolving multi-modal referring expressions in human-robot interaction. In *Proceedings of "PRE-CogSci", Workshop at 31th Annual Conference of the Cognitive Science Society*, Amsterdam, Netherlands.

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

Staudte, M. and Crocker, M. W. (2009c). Visual Attention in Spoken Human-Robot Interaction. In *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI'09)*, San Diego, USA.

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

Thórisson, K. R. (1999). A mind model for communicative humanoids. *International Journal of Applied Artificial Intelligence*, 13:449–486.

Tipples, J. (2008). Orienting to counterpredictive gaze and arrow cues. *Perception & Psychophysics*, 70:77–87.

Tomasello, M. and Carpenter, M. (2007). Shared Intentionality. *Developmental Science*, 10:121–125.

Van der Meulen, F. F., Meyer, A. S., and Levelt, W. J. M. (2001). Eye movements during the production of nouns and pronouns. *Memory & Cognition*, 29(3):512–521.

Vecera, S. and Rizzo, M. (2006). Eye gaze does not produce reflexive shifts of attention: Evidence from frontal-lobe damage. *Neuropsychologia*, 44:150–159.

Wang, E., Lignos, C., Vatsal, A., and Scassellati, B. (2006). Effects of head movement on perceptions of humanoid robot behavior. In *HRI '06: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 180–185, New York, NY, USA. ACM.

Woods, S., Walters, M., Koay, K. L., and 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*.

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

Yuki, M., Maddux, W. W., and Masuda, T. (2007). Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and the mouth as cues to recognize emotions in Japan and the United States. *Journal of Experimental Psychology*, 43:303–311.

# List of Figures

# List of Tables

# A. Item Sentences

In Experiment 1, sentences marked with a) and b) appear with scenes of the corresponding item (shown in Appendix B) that are also marked as a) and b) respectively. In Experiments 2-4, all sentences of an item appear with the same scene provided for this item. For Experiment 5, we provide an explicit mapping of sentence items with the scenes provided in Appendix B, since some items form Experiments 2-4 were discarded for this experiment.

## A.1. Experiment 1

1. a) Neben der Kugel steht eine Pyramide.

   b) Neben der Kugel steht ein Würfel.

2. a) Der Keil steht neben einem Stern.

   b) Der Keil steht neben einem Ring.

3. a) Neben dem Zylinder ist eine Halbkugel.

   b) Neben dem Zylinder ist ein Herz.

4. a) Das Ei ist neben einem Kegel.

   b) Das Ei ist neben einem Quader.

5. a) Bei dem Stern steht ein Zylinder.

   b) Bei dem Stern steht ein Keil.

6. a) Der Würfel steht bei einem Ring.

   b) Der Würfel steht bei einem Kegel.

7. a) Bei der Pyramide ist ein Ei.

   b) Bei der Pyramide ist ein Herz.

8. a) Die Halbkugel liegt bei einem Quader.
   b) Die Halbkugel liegt bei einer Kugel.
9. a) Der Quader steht neben einem Keil.
   b) Der Quader steht neben einem Ei.
10. a) Neben dem Kegel ist ein Herz.
    b) Neben dem Kegel ist eine Pyramide.
11. a) Die Pyramide ist neben einer Kugel.
    b) Die Pyramide ist neben ein Stern.
12. a) Neben der Halbkugel steht ein Würfel.
    b) Neben der Halbkugel steht ein Zylinder.
13. a) Das Herz steht bei einer Kugel.
    b) Das Herz steht bei einem Würfel.
14. a) Bei dem Zylinder liegt eine Halbkugel.
    b) Bei dem Zylinder liegt ein Quader.
15. a) Das Ei liegt bei einem Stern.
    b) Das Ei liegt bei einem Ring.
16. a) Neben der Kugel ist eine Pyramide.
    b) Neben der Kugel ist ein Keil.

## A.2. Experiments 2 and 3

1. a) Die Kugel ist größer als die Halbkugel, die hellgrün ist.
   b) Die Kugel ist größer als die Halbkugel, die gelb ist.
   c) Die Kugel ist kleiner als die Halbkugel, die hellgrün ist.
   d) Die Kugel ist kleiner als die Halbkugel, die gelb ist.
2. a) Der Zylinder ist höher als die Pyramide, die pink ist.
   b) Der Zylinder ist höher als die Pyramide, die braun ist.

c) Der Zylinder ist niedriger als die Pyramide, die pink ist.

d) Der Zylinder ist niedriger als die Pyramide, die braun ist.

3. a) Der Würfel ist breiter als die Kugel, die orange ist.

   b) Der Würfel ist breiter als die Kugel, die lila ist.

   c) Der Würfel ist schmaler als die Kugel, die orange ist.

   d) Der Würfel ist schmaler als die Kugel, die lila ist.

4. a) Die Kugel ist breiter als der Zylinder, der rosa ist.

   b) Die Kugel ist breiter als der Zylinder, der grau ist.

   c) Die Kugel ist schmaler als der Zylinder, der rosa ist.

   d) Die Kugel ist schmaler als der Zylinder, der grau ist.

5. a) Die Halbkugel ist größer als der Würfel, der rot ist.

   b) Die Halbkugel ist größer als der Würfel, der blau ist.

   c) Die Halbkugel ist kleiner als der Würfel, der rot ist.

   d) Die Halbkugel ist kleiner als der Würfel, der blau ist.

6. a) Der Quader ist breiter als das Ei, das gelb ist.

   b) Der Quader ist breiter als das Ei, das rot ist.

   c) Der Quader ist schmaler als das Ei, das gelb ist.

   d) Der Quader ist schmaler als das Ei, das rot ist.

7. a) Der Stern ist größer als der Kegel, der grün ist.

   b) Der Stern ist größer als der Kegel, der hellblau ist.

   c) Der Stern ist kleiner als der Kegel, der grün ist.

   d) Der Stern ist kleiner als der Kegel, der hellblau ist.

8. a) Der Kegel ist höher als das Herz, das silber ist.

   b) Der Kegel ist höher als das Herz, das hellgrün ist.

   c) Der Kegel ist niedriger als das Herz, das silber ist.

   d) Der Kegel ist niedriger als das Herz, das hellgrün ist.

9. a) Der Keil ist breiter als der Ring, der hellblau ist.

b) Der Keil ist breiter als der Ring, der pink ist.

c) Der Keil ist schmaler als der Ring, der hellblau ist.

d) Der Keil ist schmaler als der Ring, der pink ist.

10. a) Der Stern ist höher als der Keil, der braun ist.

b) Der Stern ist höher als der Keil, der orange ist.

c) Der Stern ist niedriger als der Keil, der braun ist.

d) Der Stern ist niedriger als der Keil, der orange ist.

11. a) Das Ei ist größer als der Quader, der schwarz ist.

b) Das Ei ist größer als der Quader, der grün ist.

c) Das Ei ist kleiner als der Quader, der schwarz ist.

d) Das Ei ist kleiner als der Quader, der grün ist.

12. a) Der Kegel ist höher als der Stern, der blau ist.

b) Der Kegel ist höher als der Stern, der schwarz ist.

c) Der Kegel ist niedriger als der Stern, der blau ist.

d) Der Kegel ist niedriger als der Stern, der schwarz ist.

13. a) Der Quader ist breiter als die Halbkugel, die hellgrün ist.

b) Der Quader ist breiter als die Halbkugel, die gelb ist.

c) Der Quader ist schmaler als die Halbkugel, die hellgrün ist.

d) Der Quader ist schmaler als die Halbkugel, die gelb ist.

14. a) Die Kugel ist gößer als die Pyramide, die pink ist.

b) Die Kugel ist gößer als die Pyramide, die braun ist.

c) Die Kugel ist kleiner als die Pyramide, die pink ist.

d) Die Kugel ist kleiner als die Pyramide, die braun ist.

15. a) Der Ring ist höher als die Kugel, die orange ist.

b) Der Ring ist höher als die Kugel, die lila ist.

c) Der Ring ist niedriger als die Kugel, die orange ist.

d) Der Ring ist niedriger als die Kugel, die lila ist.

16. a) Der Würfel ist höher als der Zylinder, der rosa ist.
    b) Der Würfel ist höher als der Zylinder, der grau ist.
    c) Der Würfel ist niedriger als der Zylinder, der rosa ist.
    d) Der Würfel ist niedriger als der Zylinder, der grau ist.

17. a) Die Pyramide ist breiter als der Würfel, der rot ist.
    b) Die Pyramide ist breiter als der Würfel, der blau ist.
    c) Die Pyramide ist schmaler als der Würfel, der rot ist.
    d) Die Pyramide ist schmaler als der Würfel, der blau ist.

18. a) Die Halbkugel ist größer als das Ei, das gelb ist.
    b) Die Halbkugel ist größer als das Ei, das rot ist.
    c) Die Halbkugel ist kleiner als das Ei, das gelb ist.
    d) Die Halbkugel ist kleiner als das Ei, das rot ist.

19. a) Das Herz ist höher als der Kegel, der grün ist.
    b) Das Herz ist höher als der Kegel, der hellblau ist.
    c) Das Herz ist niedriger als der Kegel, der grün ist.
    d) Das Herz ist niedriger als der Kegel, der hellblau ist.

20. a) Der Zylinder ist größer als das Herz, das silber ist.
    b) Der Zylinder ist größer als das Herz, das hellgrün ist.
    c) Der Zylinder ist kleiner als das Herz, das silber ist.
    d) Der Zylinder ist kleiner als das Herz, das hellgrün ist.

21. a) Die Pyramide ist höher als der Ring, der hellblau ist.
    b) Die Pyramide ist höher als der Ring, der pink ist.
    c) Die Pyramide ist niedriger als der Ring, der hellblau ist.
    d) Die Pyramide ist niedriger als der Ring, der pink ist.

22. a) Der Würfel ist breiter als der Keil, der braun ist.
    b) Der Würfel ist breiter als der Keil, der orange ist.
    c) Der Würfel ist schmaler als der Keil, der braun ist.

d) Der Würfel ist schmaler als der Keil, der orange ist.

23. a) Der Ring ist breiter als der Quader, der schwarz ist.
 b) Der Ring ist breiter als der Quader, der grün ist.
 c) Der Ring ist schmaler als der Quader, der schwarz ist.
 d) Der Ring ist schmaler als der Quader, der grün ist.

24. a) Das Herz ist größer als der Stern, der blau ist.
 b) Das Herz ist größer als der Stern, der schwarz ist.
 c) Das Herz ist kleiner als der Stern, der blau ist.
 d) Das Herz ist kleiner als der Stern, der schwarz ist.

## A.3. Experiment 4

1. a) Die Kugel ist größer als die grüne Halbkugel.
 b) Die grüne Halbkugel ist kleiner als die Kugel.
 c) Die Kugel ist kleiner als die gelbe Halbkugel.
 d) Die gelbe Halbkugel ist größer als die Kugel.

2. a) Der Zylinder ist höher als die pinke Pyramide.
 b) Die pinke Pyramide ist niedriger als der Zylinder.
 c) Der Zylinder is niedriger als die braune Pyramide.
 d) Die braune Pyramide ist höher als der Zylinder.

3. a) Der Würfel ist breiter als die orange Kugel.
 b) Die orange Kugel ist schmaler als der Würfel.
 c) Der Würfel ist schmaler als die lila Kugel.
 d) Die lila Kugel ist breiter als der Würfel.

4. a) Die Kugel ist breiter als der rosa Zylinder.
 b) Der rosa Zylinder ist schmaler als die Kugel.
 c) Die Kugel ist schmaler als der graue Zylinder.
 d) Der graue Zylinder ist breiter als die Kugel.

5. a) Die Halbkugel ist größer als der rote Würfel.
   b) Der rote Würfel ist kleiner als die Halbkugel.
   c) Die Halbkugel ist kleiner als der blaue Würfel.
   d) Der blaue Würfel ist größer als die Halbkugel.

6. a) Der Quader ist breiter als das gelbe Ei.
   b) Das gelbe Ei ist schmaler als der Quader.
   c) Der Quader ist schmaler als das rote Ei.
   d) Das rote Ei ist breiter als der Quader.

7. a) Der Stern ist größer als der grüne Kegel.
   b) Der grüne Kegel ist kleiner als der Stern.
   c) Der Stern ist kleiner als der blaue Kegel.
   d) Der blaue Kegel ist größer als der Stern.

8. a) Der Kegel ist höher als das silberne Herz.
   b) Das silberne Herz ist niedriger als der Kegel.
   c) Der Kegel ist niedriger als das grüne Herz.
   d) Das grüne Herz ist höher als der Kegel.

9. a) Der Keil ist breiter als der blaue Ring.
   b) Der blaue Ring ist schmaler als der Keil.
   c) Der Keil ist schmaler als der pinke Ring.
   d) Der pinke Ring ist breiter als der Keil.

10. a) Der Stern ist höher als der braune Keil.
    b) Der braune Keil ist niedriger als der Stern.
    c) Der Stern ist niedriger als der orange Keil.
    d) Der orange Keil ist höher als der Stern.

11. a) Das Ei ist größer als der schwarze Quader.
    b) Der schwarze Quader ist kleiner als das Ei.
    c) Das Ei ist kleiner als der grüne Quader.

d) Der grüne Quader ist größer als das Ei.

12. a) Der Kegel ist höher als der blaue Stern.

    b) Der blaue Stern ist niedriger als der Kegel.

    c) Der Kegel ist niedriger als der schwarze Stern.

    d) Der schwarze Stern ist höher als der Kegel.

13. a) Der Quader ist breiter als die grüne Halbkugel,.

    b) Die grüne Halbkugel ist schmaler als der Quader.

    c) Der Quader ist schmaler als die gelbe Halbkugel.

    d) Die gelbe Halbkugel ist breiter als der Quader.

14. a) Die Kugel ist gößer als die pinke Pyramide.

    b) Die pinke Pyramide ist kleiner als die Kugel.

    c) Die Kugel ist kleiner als die braune Pyramide.

    d) Die braune Pyramider ist größer als die Kugel.

15. a) Der Ring ist höher als die orange Kugel.

    b) Die orange Kugel ist niedriger als der Ring.

    c) Der Ring ist niedriger als die lila Kugel.

    d) Die lila Kugel ist höher als der Ring.

16. a) Der Würfel ist höher als der rosa Zylinder.

    b) Der rosa Zylinder ist niedriger als der Würfel.

    c) Der Würfel ist niedriger als der graue Zylinder.

    d) Der graue Zylinder ist höher als der Würfel.

17. a) Die Pyramide ist breiter als der rote Würfel.

    b) Der rote Würfel ist schmaler als die Pyramide.

    c) Die Pyramide ist schmaler als der blaue Würfel.

    d) Der blaue Würfel ist breiter als die Pyramide.

18. a) Die Halbkugel ist größer als das gelbe Ei.

    b) Das gelber Ei ist kleiner als die Halbkugel.

c) Die Halbkugel ist kleiner als das rote Ei.

d) Das rote Ei ist größer als die Halbkugel.

19. a) Das Herz ist höher als der grüne Kegel.

b) Der grüne Kegel ist niedriger als das Herz.

c) Das Herz ist niedriger als der blaue Kegel.

d) Der blaue Kegel ist höher als das Herz.

20. a) Der Zylinder ist größer als das silberne Herz.

b) Das silberne Herz ist kleiner als der Zylinder.

c) Der Zylinder ist kleiner als das grüne Herz.

d) Das grüne Herz ist größer als der Zylinder.

21. a) Die Pyramide ist höher als der blaue Ring.

b) Der blaue Ring ist niedriger als die Pyramide.

c) Die Pyramide ist niedriger als der pinke Ring.

d) Der pinke Ring ist höher als die Pyramide.

22. a) Der Würfel ist breiter als der braune Keil.

b) Der braune Keil ist schmaler als der Würfel.

c) Der Würfel ist schmaler als der orange Keil.

d) Der orange Keil ist breiter als der Würfel.

23. a) Der Ring ist breiter als der schwarze Quader.

b) Der schwarze Quader ist schmaler als der Ring.

c) Der Ring ist schmaler als der grüne Quader.

d) Der grüne Quader ist breiter als der Ring.

24. a) Das Herz ist größer als der blaue Stern.

b) Der blaue Stern ist kleiner als das Herz.

c) Das Herz ist kleiner als der schwarze Stern.

d) Der schwarze Stern ist größer als das Herz.

## A.4. Experiment 5

1. (Scene 1)
   a) Die rote Kugel ist größer als die grüne Halbkugel.
   b) Die grüne Halbkugel ist kleiner als die rote Kugel.
   c) Die rote Kugel ist kleiner als die gelbe Halbkugel.
   d) Die gelbe Halbkugel ist größer als die rote Kugel.

2. (Scene 2)
   a) Der orange Zylinder ist höher als die pinke Pyramide.
   b) Die pinke Pyramide ist niedriger als der orange Zylinder.
   c) Der orange Zylinder is niedriger als die braune Pyramide.
   d) Die braune Pyramide ist höher als der orange Zylinder.

3. (Scene 6)
   a) Der blaue Quader ist breiter als das gelbe Ei.
   b) Das gelbe Ei ist schmaler als der blaue Quader.
   c) Der blaue Quader ist schmaler als das rote Ei.
   d) Das rote Ei ist breiter als der blaue Quader.

4. (Scene 7)
   a) Der schwarze Stern ist größer als der grüne Kegel.
   b) Der grüne Kegel ist kleiner als der schwarze Stern.
   c) Der schwarze Stern ist kleiner als der blaue Kegel.
   d) Der blaue Kegel ist größer als der schwarze Stern.

5. (Scene 8)
   a) Der gelbe Kegel ist höher als das silberne Herz.
   b) Das silberne Herz ist niedriger als der gelbe Kegel.
   c) Der gelbe Kegel ist niedriger als das grüne Herz.
   d) Das grüne Herz ist höher als der gelbe Kegel.

6. (Scene 9)

a) Der graue Keil ist breiter als der blaue Ring.

b) Der blaue Ring ist schmaler als der graue Keil.

c) Der graue Keil ist schmaler als der pinke Ring.

d) Der pinke Ring ist breiter als der graue Keil.

7. (Scene 10)

a) Der blaue Stern ist höher als der braune Keil.

b) Der braune Keil ist niedriger als der blaue Stern.

c) Der blaue Stern ist niedriger als der orange Keil.

d) Der orange Keil ist höher als der blaue Stern.

8. (Scene 11)

a) Das pinke Ei ist größer als der schwarze Quader.

b) Der schwarze Quader ist kleiner als das pinke Ei.

c) Das pinke Ei ist kleiner als der grüne Quader.

d) Der grüne Quader ist größer als das pinke Ei.

9. (Scene 13)

a) Der blaue Quader ist breiter als die grüne Halbkugel,.

b) Die grüne Halbkugel ist schmaler als der blaue Quader.

c) Der blaue Quader ist schmaler als die gelbe Halbkugel.

d) Die gelbe Halbkugel ist breiter als der blaue Quader.

10. (Scene 14)

a) Die rote Kugel ist gößer als die pinke Pyramide.

b) Die pinke Pyramide ist kleiner als die rote Kugel.

c) Die rote Kugel ist kleiner als die braune Pyramide.

d) Die braune Pyramider ist größer als die rote Kugel.

11. (Scene 15)

a) Der braune Ring ist höher als die orange Kugel.

b) Die orange Kugel ist niedriger als der braune Ring.

c) Der braune Ring ist niedriger als die lila Kugel.

d) Die lila Kugel ist höher als der braune Ring.

12. (Scene 16)

a) Der grüne Würfel ist höher als der rosa Zylinder.

b) Der rosa Zylinder ist niedriger als der grüne Würfel.

c) Der grüne Würfel ist niedriger als der graue Zylinder.

d) Der graue Zylinder ist höher als der grüne Würfel.

13. (Scene 17)

a) Die grüne Pyramide ist breiter als der rote Würfel.

b) Der rote Würfel ist schmaler als die grüne Pyramide.

c) Die grüne Pyramide ist schmaler als der blaue Würfel.

d) Der blaue Würfel ist breiter als die grüne Pyramide.

14. (Scene 18)

a) Die blaue Halbkugel ist größer als das gelbe Ei.

b) Das gelber Ei ist kleiner als die blaue Halbkugel.

c) Die blaue Halbkugel ist kleiner als das rote Ei.

d) Das rote Ei ist größer als die blaue Halbkugel.

15. (Scene 19)

a) Das lila Herz ist höher als der grüne Kegel.

b) Der grüne Kegel ist niedriger als das lila Herz.

c) Das lila Herz ist niedriger als der blaue Kegel.

d) Der blaue Kegel ist höher als das lila Herz.

16. (Scene 20)

a) Der orange Zylinder ist größer als das silberne Herz.

b) Das silberne Herz ist kleiner als der orange Zylinder.

c) Der orange Zylinder ist kleiner als das grüne Herz.

d) Das grüne Herz ist größer als der orange Zylinder.

17. (Scene 22)
    a) Der grüne Würfel ist breiter als der braune Keil.
    b) Der braune Keil ist schmaler als der grüne Würfel.
    c) Der grüne Würfel ist schmaler als der orange Keil.
    d) Der orange Keil ist breiter als der grüne Würfel.

18. (Scene 23)
    a) Der braune Ring ist breiter als der schwarze Quader.
    b) Der schwarze Quader ist schmaler als der braune Ring.
    c) Der braune Ring ist schmaler als der grüne Quader.
    d) Der grüne Quader ist breiter als der braune Ring.

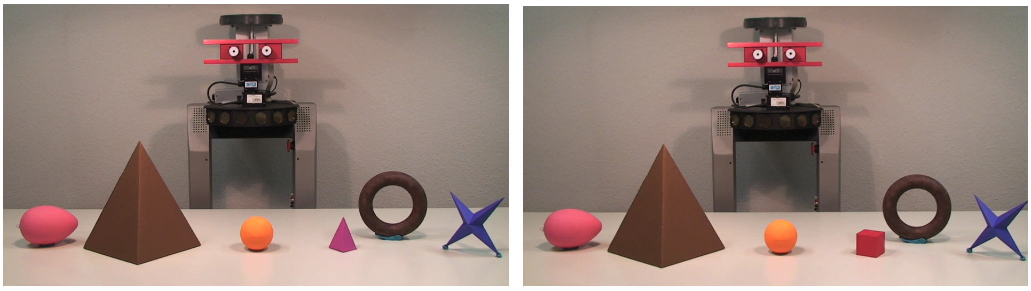
19. (Scene 24)
    a) Das lila Herz ist größer als der blaue Stern.
    b) Der blaue Stern ist kleiner als das lila Herz.
    c) Das lila Herz ist kleiner als der schwarze Stern.
    d) Der schwarze Stern ist größer als das lila Herz.

20. (Scene 4)
    a) Die rote Kugel ist breiter als der rosa Zylinder.
    b) Der rosa Zylinder ist schmaler als die rote Kugel.
    c) Die rote Kugel ist schmaler als der graue Zylinder.
    d) Der graue Zylinder ist breiter als die rote Kugel.

# B. Item Scenes

## B.1. Experiment 1

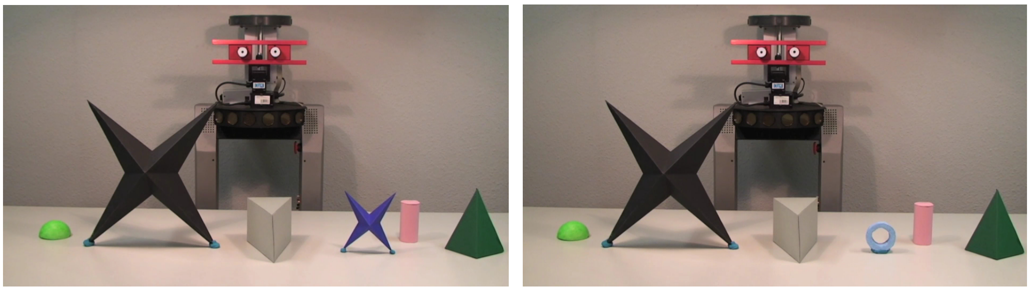
### Item 1



(a) ambiguous

(b) unambiguous

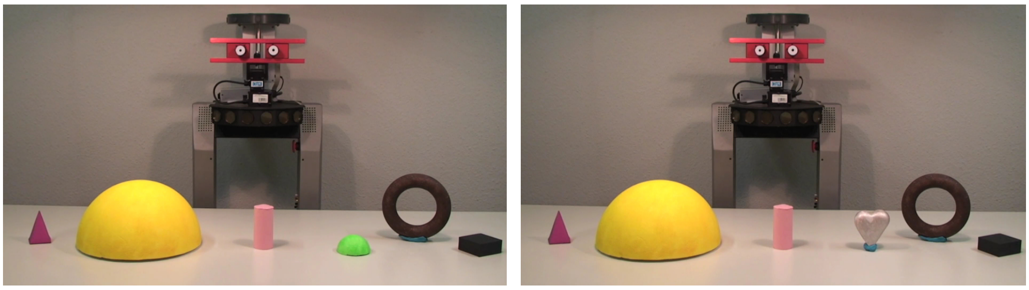
### Item 2



(a) ambiguous

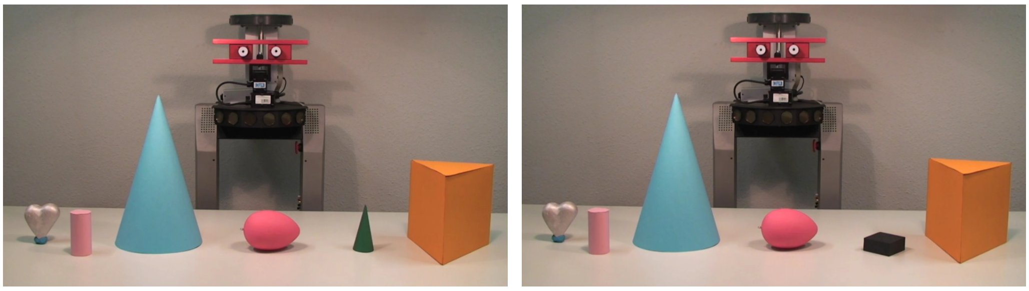
(b) unambiguous

### Item 3



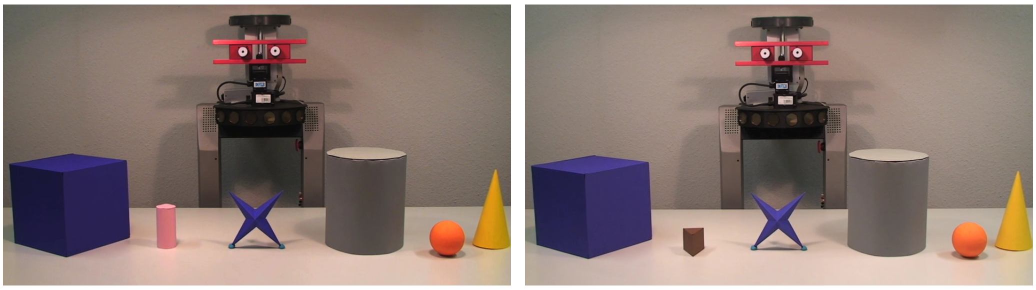
(a) ambiguous (b) unambiguous

### Item 4



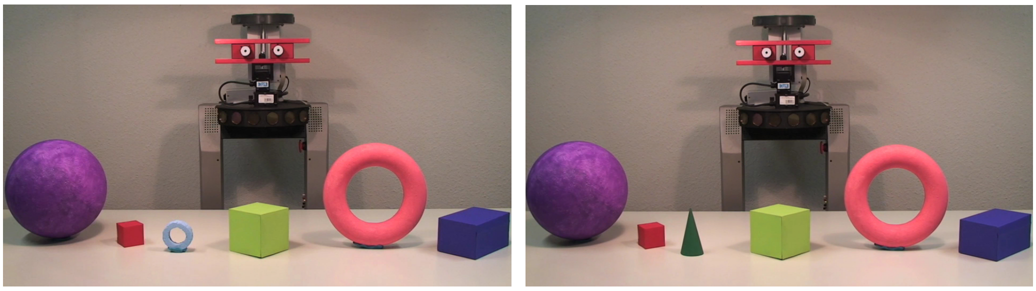
(a) ambiguous (b) unambiguous

### Item 5



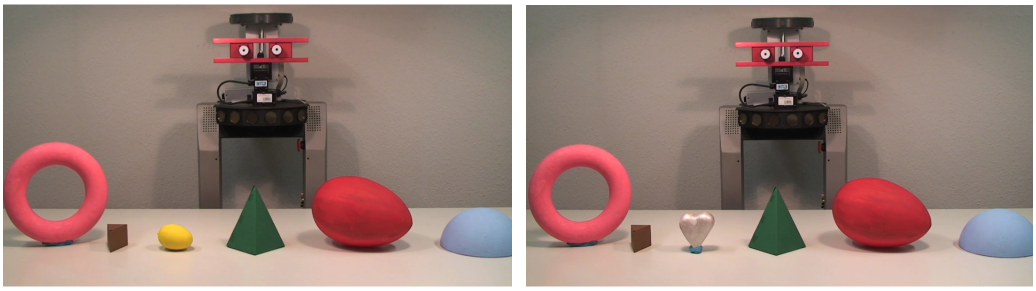
(a) ambiguous (b) unambiguous

### Item 6



(a) ambiguous (b) unambiguous

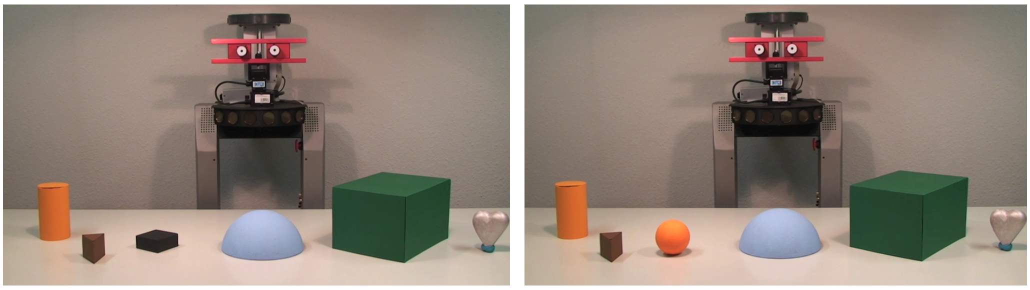
### Item 7



(a) ambiguous

(b) unambiguous

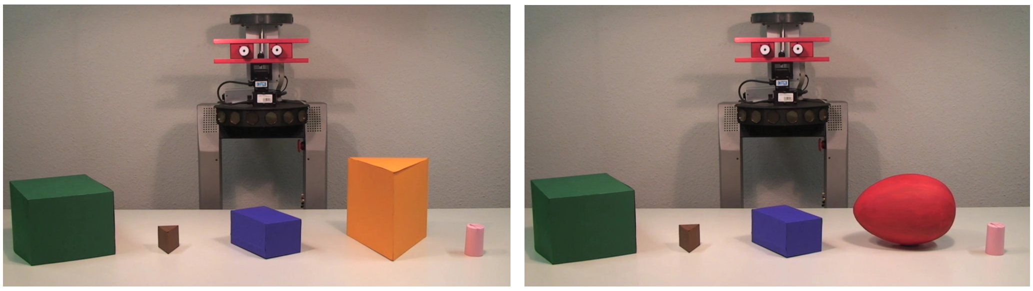
### Item 8



(a) ambiguous

(b) unambiguous

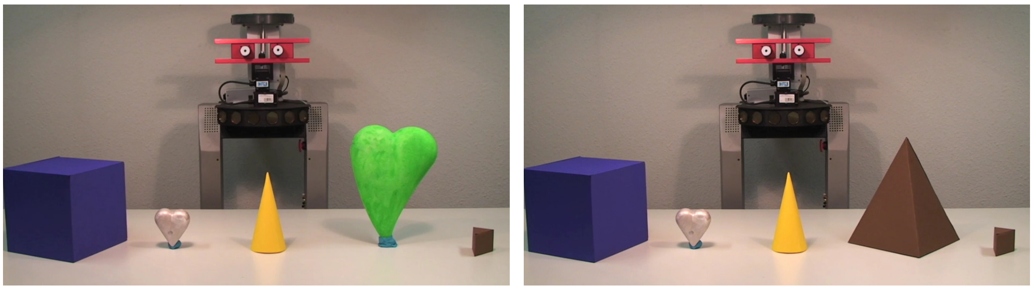
### Item 9



(a) ambiguous

(b) unambiguous

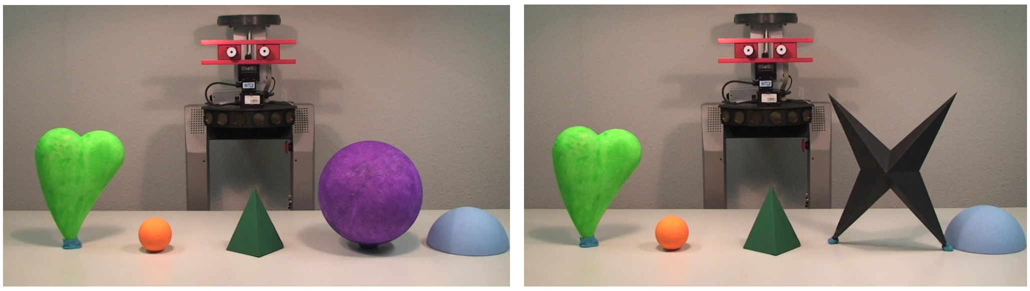
### Item 10



(a) ambiguous

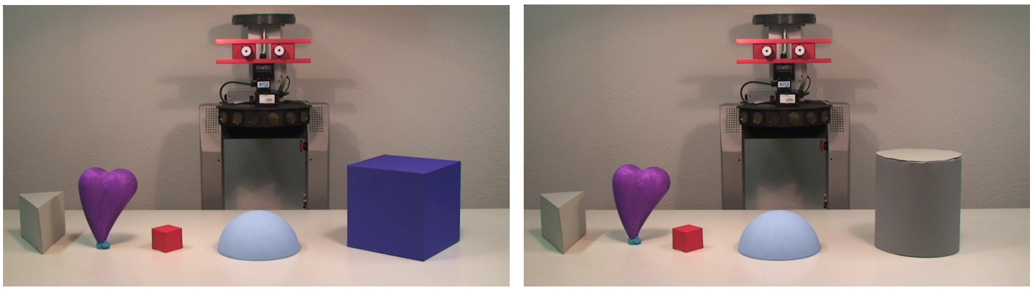
(b) unambiguous

### Item 11



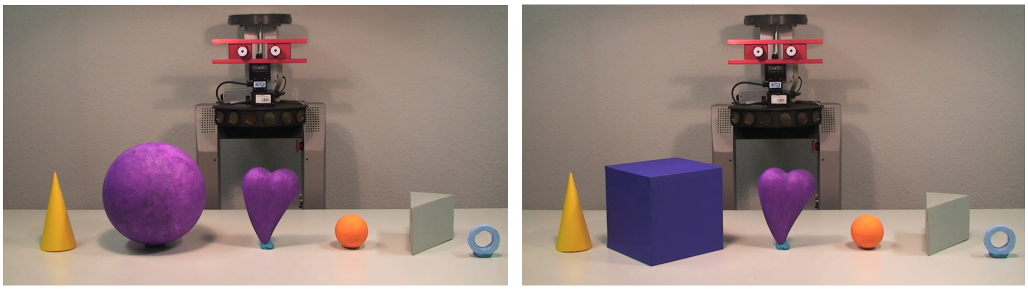
(a) ambiguous (b) unambiguous

### Item 12



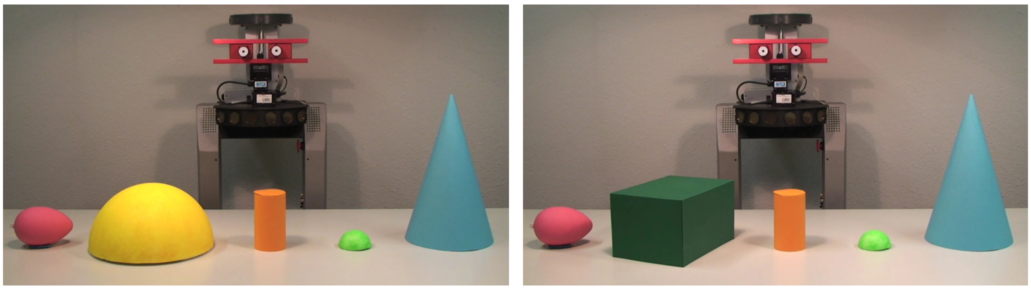
(a) ambiguous (b) unambiguous

### Item 13



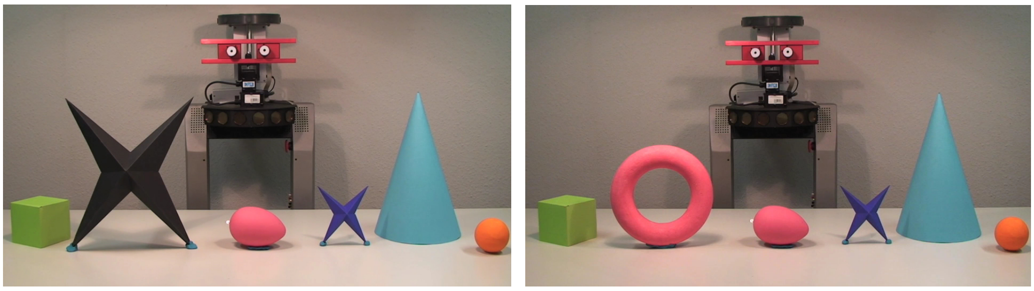
(a) ambiguous (b) unambiguous

### Item 14



(a) ambiguous (b) unambiguous

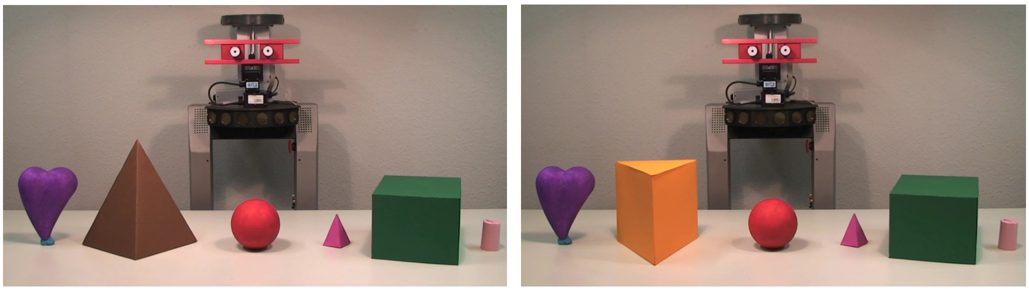
## Item 15



(a) ambiguous

(b) unambiguous

## Item 16

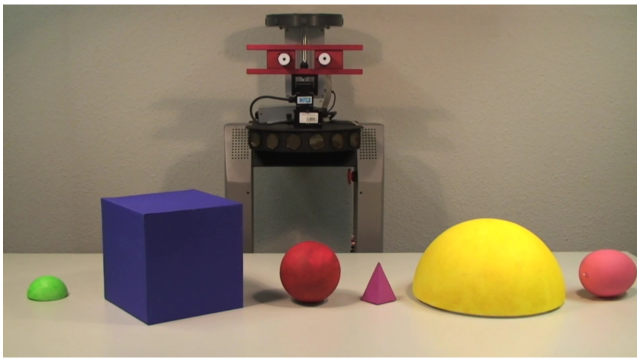


(c) ambiguous

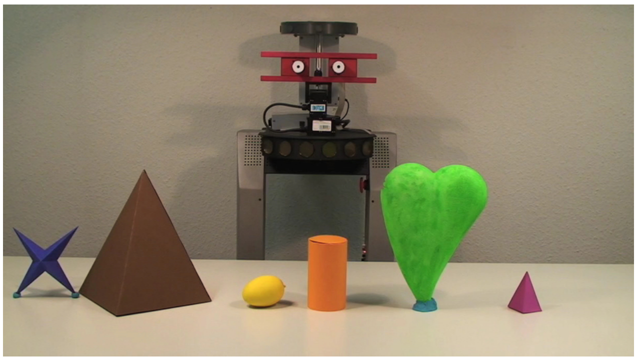
(d) unambiguous

## B.2. Experiments 2 - 5

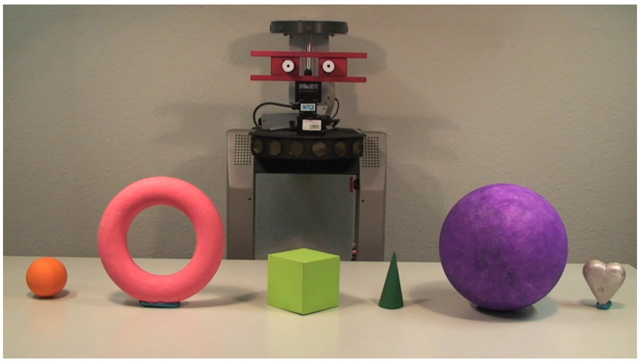
**Item 1**



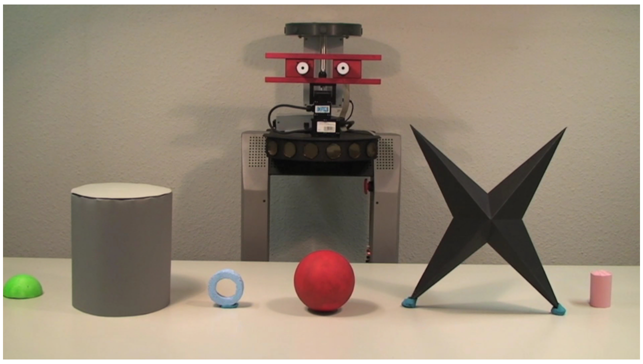
**Item 2**



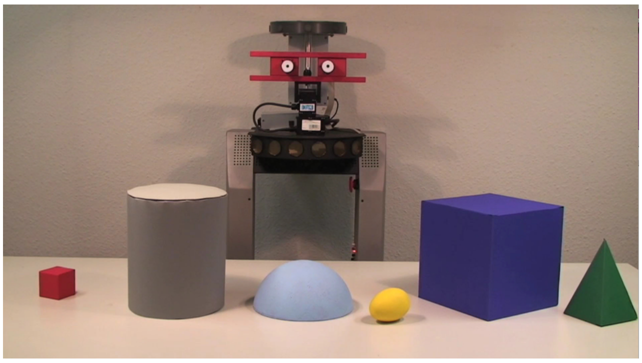
**Item 3**



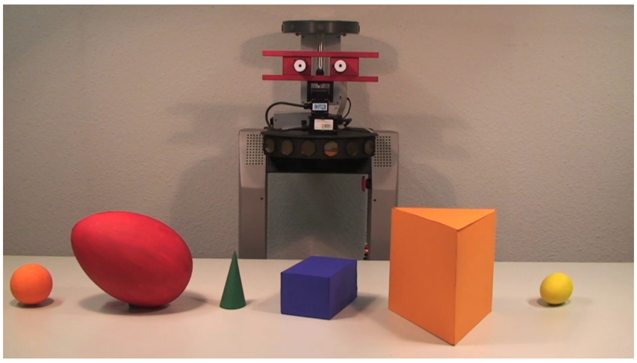
**Item 4**



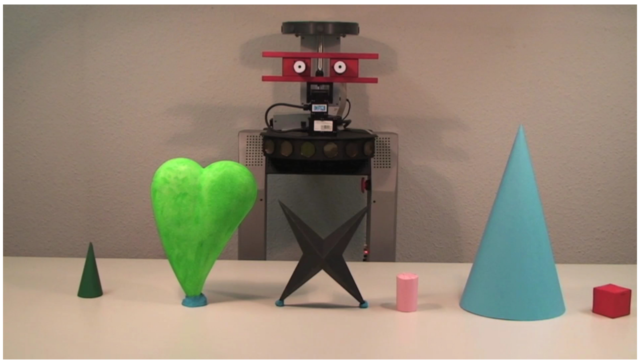
**Item 5**



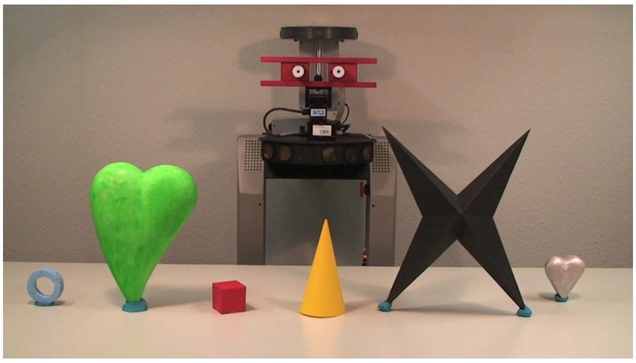
Item 6



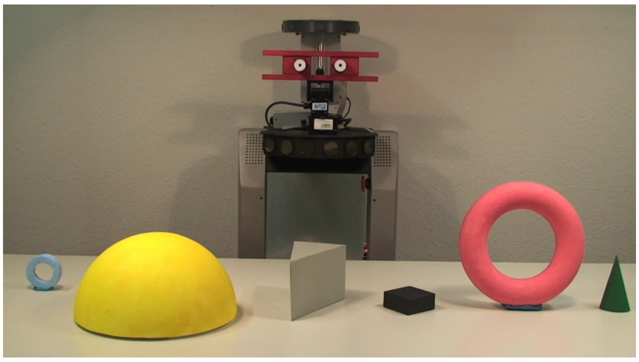
Item 7



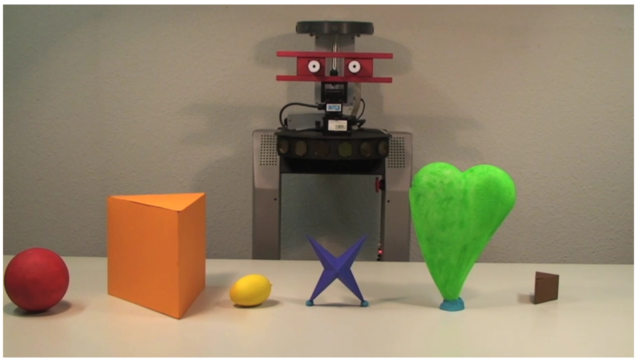
Item 8



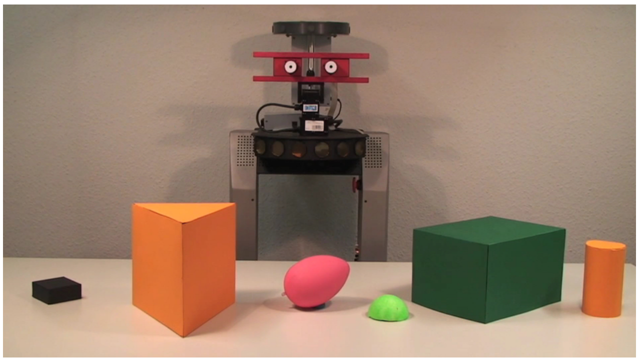
**Item 9**



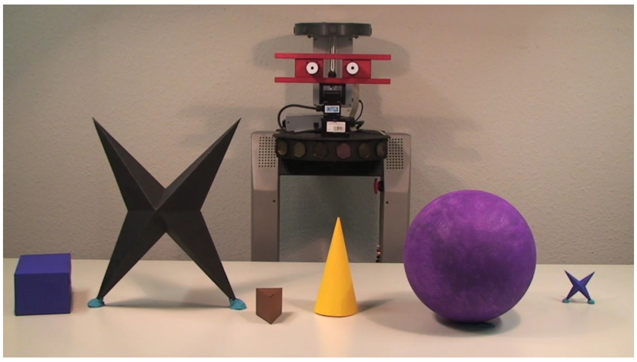
**Item 10**



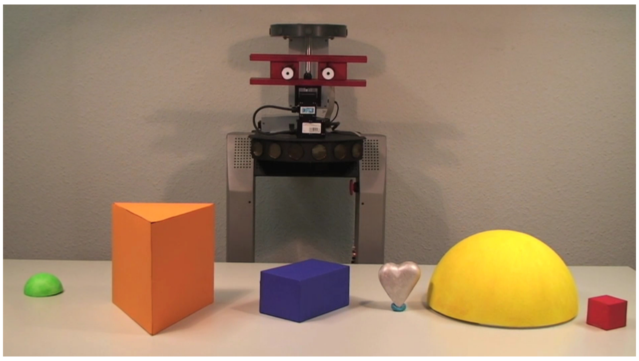
**Item 11**



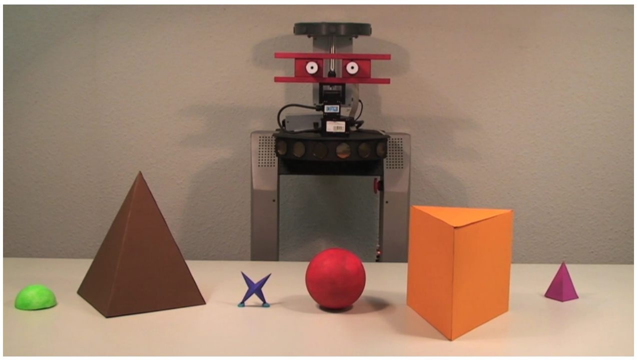
**Item 12**



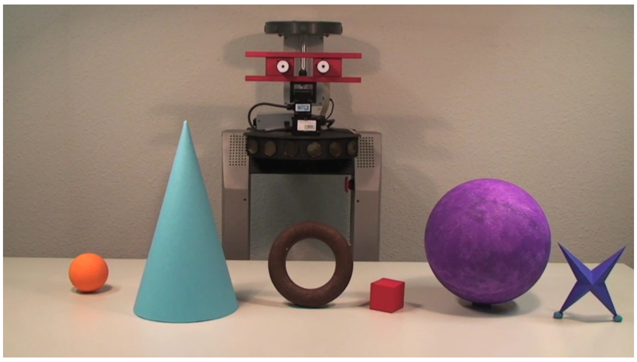
**Item 13**



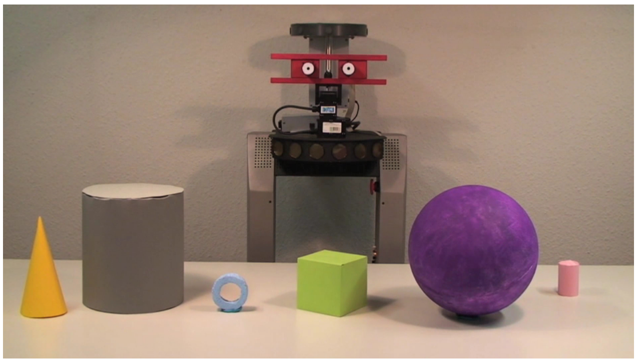
**Item 14**



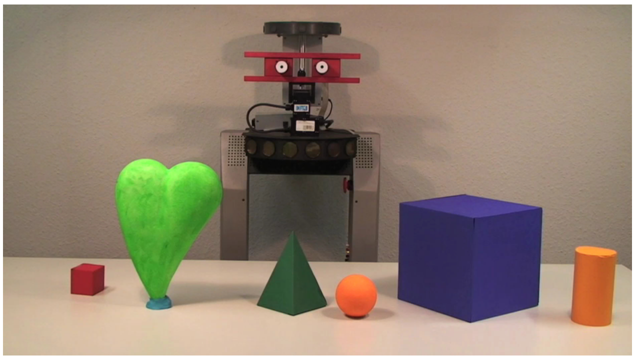
**Item 15**



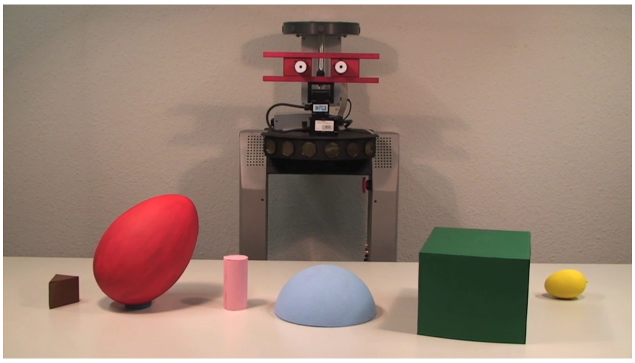
**Item 16**



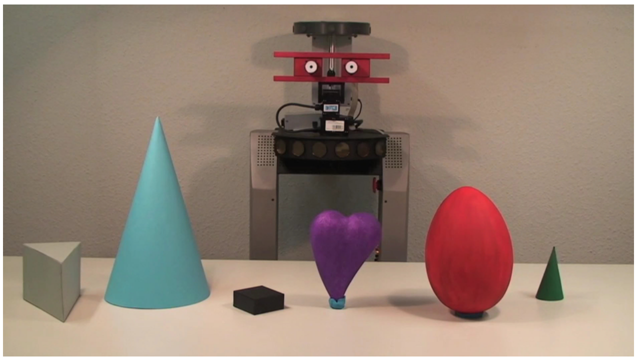
**Item 17**



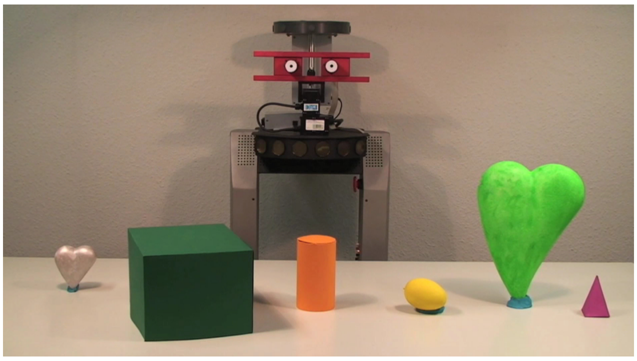
**Item 18**



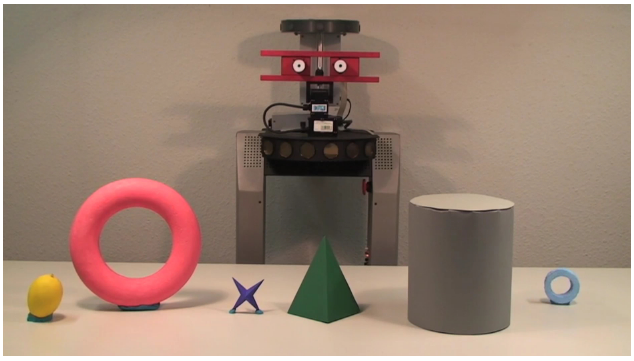
**Item 19**



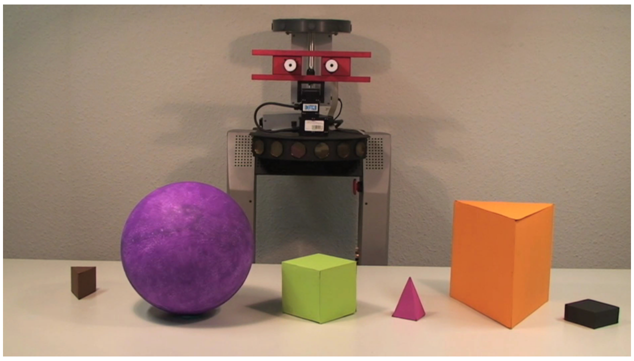
**Item 20**



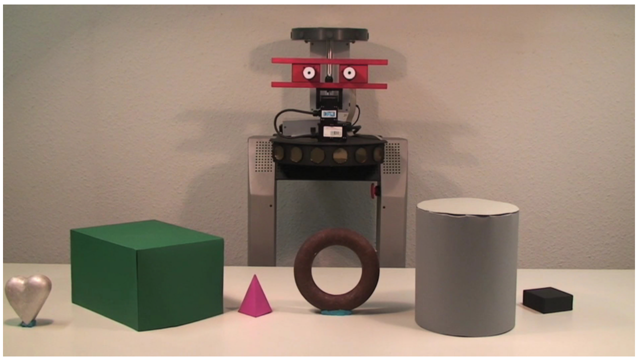
**Item 21**



**Item 22**



**Item 23**



### Item 24