Reference-Related Speaker Gaze as a Cue in Online Sentence Processing

Helene Kreysa [*] **Pia Knoeferle** [*]

[*] *Cognitive Interaction Technology Excellence Cluster (CITEC), Universität Bielefeld, Morgenbreede 39, 33615 Bielefeld (e-mail: helene.kreysa@uni-jena.de)*

Abstract: We report a series of eye-tracking studies investigating different facets of how seeing a speaker's gaze affects listeners' visual attention and comprehension. We compare the effect of speaker gaze to other cues in the linguistic and non-linguistic context, such as depicted actions and sentence structure. In addition, we discuss top-down influences of the comprehension sub-task, as well as the similarities and differences between using the gaze of a human speaker and that of an artificial agent. Our results suggest that human listeners rapidly make use of speaker gaze as a cue to upcoming conversational content, and that this benefit generalises across a range of situations. At the same time, the extent of gaze-following is affected by sentence structure. These findings are important for processing accounts and models of situated language comprehension, but they also contribute to developing communicative agents that behave in a natural and human-like way.

Keywords: situated language processing, spoken sentence comprehension, referential gaze, virtual agent

1. INTRODUCTION

Attending to the same aspects of the visual surroundings as your conversational partner can improve your understanding of what he or she is saying (e.g., Richardson & Dale, 2005; Richardson et al., 2007). One way that such coordination of attention can come about is by looking in the same direction as your conversational partner. Indeed, the ability to follow a speaker's gaze has long been known to be helpful for comprehension (see e.g., Kleinke, 1986). In part, this is because human speakers tend to look at objects they are about to mention (Meyer et al., 1998; Griffin & Bock, 2000). This means that their gaze can allow referential disambiguation, often even before the object's name is mentioned (Hanna & Brennan, 2007; Staudte & Crocker, 2011). In fact, this systematic link between looking at an object in the world and referring to it may be one reason for people's tendency to attend to objects in the line of other people's gaze (Driver et al., 1999; Ricciardelli et al., 2002; Castelhano et al., 2007). This association arguably also motivates the key role of speaker gaze in disambiguating the mapping between objects and words in child language acquisition (Baldwin, 1993; 1995; see also Yu et al., 2005). Yet despite a general consensus that gaze is a helpful cue in conversation, little is known to date about the extent to which referential disambiguation through speaker gaze generalises across different sentence structures, tasks, and positions of the speaker relative to the listener.

To the extent that speaker gaze affects language comprehension rapidly and robustly, theories and accounts of language comprehension should accommodate its effects. Eye-tracking research on language comprehension has shown that listeners rapidly integrate information from the visual context with the unfolding speech content (e.g., Tanenhaus et al., 1995; Sedivy et al., 1999; Spivey et al., 2001; Kamide et al., 2003; Altmann, 2004; Knoeferle et al., 2005). Given such early influences of the visual world, the speaker's direction of visual attention may be processed similarly rapidly, as an integral component of visual context. If this is the case, details of its integration with the speech stream can and should be accommodated by existing accounts of situated comprehension (e.g., Knoeferle & Crocker, 2006; 2007; Altmann & Kamide, 2007) and associated computational models (e.g., Mayberry et al., 2009; Crocker et al., 2010).

From a practical perspective, an accurate model of how speaker gaze is used by a comprehender or user would also be desirable, since speech-related gaze behaviour may prove to be a key ingredient in designing communicative agents that behave in a natural and human-like fashion when interacting with a human user (e.g., Kopp, 2010). In the following sections, we discuss the theoretical and practical implications of our research on speaker gaze effects, before turning to the empirical data.

1.1 The Coordinated Interplay Account of situated sentence comprehension

Eye-tracking has been used extensively in recent years to investigate the online comprehension of spoken language. The methodology is particularly suited to this purpose because eye movements are easy-to-measure indicators of visual attention: Any shift of fixation to a stimulus is preceded by a shift of attention to that object (cf. Hoffmann & Subramaniam, 1995; Deubel & Schneider, 1996). Psycholinguistic experiments in the tradition of the so-called "visual world paradigm" record human participants' eye movements while they listen to a sentence containing a temporary linguistic ambiguity (Tanenhaus et al., 1995). Simultaneously, they see either an arrangement of real-world objects in front of them, or a computer display of semi-realistic clipart objects and depicted actions. These displays are selected in such a way that a rapid integration of the information from the visual world would enable participants to disambiguate between potential sentence meanings during

real-time comprehension. In this context, looks to an object are interpreted as reflecting that a listener has adopted a specific interpretation of the grammatical function and thematic role of that object in the unfolding sentence.

For example, consider German object-verb-subject (OVS) sentences like (1), versus subject-verb-object (SVO) sentences (2):

(1) Die Prinzessin malt offensichtlich der Fechter$_{Nom}$ (roughly: 'the princess is being painted by the fencer').
(2) Die Prinzessin wäscht offensichtlich den Pirat$_{Acc}$ ('the princess is washing the pirate').

Both structures are grammatical, but subject-initial sentences are canonical, while object-initial sentences are much less frequent and harder to process. Knoeferle and colleagues (2005) presented such spoken sentences to participants who saw an image of a princess, a pirate, and a fencer. The princess held a bucket and sponge in her hand, suggesting that she was a suitable agent of a "washing" action, while the fencer held a paintbrush and paints. *Prinzessin* is ambiguously case-marked; hence at this point in the sentence it is unclear whether it functions as the grammatical subject or object, and whether its thematic role is agent or patient, respectively. Despite this temporary structural ambiguity, participants looked more at the pirate during both verbs, as if they expected a canonical SVO sentence. However, right after hearing *malt*, they began looking more at the fencer (the depicted agent of the painting action). This suggests that they used the visually available information to assign an agent role to the fencer and correspondingly a patient role to the initially ambiguous princess, thus updating their understanding and thematic role assignment of the sentence.

These results – as well as similar disambiguating effects of referential visual contexts (e.g., Tanenhaus et al., 1995), case marking and world knowledge (e.g., Kamide, Scheepers, & Altmann, 2003), and even intonation (e.g., Weber et al., 2006) – have informed the so-called Coordinated Interplay Account (CIA) of situated utterance comprehension (Knoeferle & Crocker, 2006; 2007; Crocker et al., 2010). It describes how linguistic processing of the utterance guides attention to relevant scene information and/or corresponding working memory representations. In turn, attended aspects of the scene can affect how the linguistic utterance is interpreted, creating a tight temporal link between the assignment of reference and visual interrogation of the scene.

The model assumes three steps in processing each content word$_i$ of an unfolding sentence (these steps can occur in parallel, but are informationally dependent upon each other): In the first step, Incremental Sentence Interpretation, word$_i$ is integrated with the interpretation of the sentence so far, and expectations are derived for its continuation. In the subsequent step of Language-Mediated Attention, this interpretation and the associated expectations guide attention to suitable referents in the visual world or to working memory representations thereof. Once such referents are located, Scene Integration updates the linguistic interpretation with information from the scene. For instance, the mapping of the word *malt* to the paintbrush held by the fencer clarifies that the princess is not the agent but the patient of the painting action. Thus, the depicted event is used to assign the correct thematic roles, leading to a revision of the predicted sentence structure from SVO to OVS (Knoeferle et al., 2008). The following words (e.g., *der Fechter*) are then processed in the same stages, but based on the updated sentence interpretation, which now specifies thematic role relations. This interaction of scene and sentence can even allow participants to launch anticipatory eye movements to upcoming referents, before they hear the corresponding referential expression.

A computational implementation of the CIA is provided by CIANet (Mayberry et al., 2009; Crocker et al., 2010), a connectionist model based on a simple recurrent network that was extended with a representation of two scene events. The model is trained to incrementally assign case role representations to the linguistic input, based on learned syntactic and lexical constraints for anticipating likely role fillers. Processing of each new input word in the context of the currently available sentence interpretation – as described above – is realised by a copy of the previous state of the hidden layer in the simple recurrent network, which provides contextual input to the current representation. Context effects are moderated by an attentional gating vector, which increases the salience of the most relevant event at the expense of the alternative event (for more detail, see Mayberry et al., 2009). In this way, the most relevant component of the scene is selected as the target of attention, paralleling empirical fixation patterns and even neurobehavioral findings (Knoeferle et al., 2008; Crocker et al., 2010). Importantly, this performance is achieved through an incremental comprehension mechanism interacting with an explicit attentional component. CIANet uses all relevant information as soon as it becomes available, leading to adaptive and (in this regard) cognitively plausible model behaviour (Mayberry et al., 2009).

1.2 Gaze behaviour in communicative agents

In typical *visual-world* studies, the spoken sentence tends to be pre-recorded and the speaker is not visible. As a consequence, potential "interlocutor" cues, such as eye gaze, gestures, and facial expressions, are not accommodated by current accounts of spoken sentence comprehension, although there is ample reason to believe that such behaviours form an important part of the context in which utterances are produced and comprehended (e.g., Kendon, 1967; Clark, 1996; Bavelas & Chovil, 2000). With regard to gaze, humans show a strong tendency to follow the gaze of other people – indeed, gaze-cueing seems to be automatic in the sense that it is unintentional and insensitive to cognitive load (Xu et al., 2011). Consequently, it seems likely that a speaker's gaze could play a key role in directing comprehenders' attention to relevant (linguistic and/or scene) information when processing spoken sentences.

If this is the case, providing communicative agents with human-like gaze behaviour could be important in designing agents that communicate effectively (i.e., direct attention to what is relevant) and are accepted by their human

interlocutors (see Kopp, 2010). For this reason, our research has not only focused on examining effects of a human speaker's gaze but also transports an experimental paradigm with a human speaker to a context where the speaker is an artificial agent (Kreysa, Knoeferle, Yaghoubzadeh, & Kopp, in progress). On the one hand, replicating our "human" results with an agent will allow us to formulate recommendations on how to design artificial agents that employ gaze in a human-like way. Human-like behaviour has been shown to increase acceptance of artificial agents and to positively influence interactions with them (e.g., Bailenson & Yee, 2005; Breazeal et al., 2005; Sidner et al., 2005; Krach, et al. 2008; Bergmann et al., 2010; Meltzoff et al., 2010). On the other hand, we are interested in exploring the validity of gaze, and the extent to which it can cue sentence content. Ultimately, this will require experiments where the speaker looks at objects s/he is not talking about, or where his/her timing is not realistic. Virtual speakers may provide a useful tool for manipulating these aspects of speaker gaze, since their behaviour can be easily scripted (see Staudte & Crocker, 2011).

1.3 Overview of the experiments

The present research program investigates many of the issues discussed so far. It comprises seven experiments that manipulate different aspects which might affect how listeners process a speaker's gaze. All experiments used a variant of the visual-world paradigm, in which participants see a scene depicting three characters (see Fig. 1: a waiter in the middle; a millionaire to the right, and a saxophone player on the left) and hear a related transitive sentence (e.g., *the waiter congratulates the millionaire*). In half of the trials, they could see the speaker of the sentence looking at the scene (Fig.1a), while in the other half the speaker was occluded (Fig.1b).

Fig. 1: Example of a still from the video stimuli, either (a) showing the speaker or (b) with the speaker obscured.

In all experiments, participants were asked to verify whether a schematic depiction of the sentence correctly represented a specified aspect of the event described. We manipulated the comprehension task by varying what aspect this was: Participants verified either the two mentioned referents (e.g., true: the waiter and the millionaire; false: the saxophone player), the patient of the sentence (the millionaire), or the thematic role relations (waiter: agent, millionaire: patient, action goes from central position to right side of screen). The verification task was presented after the end of the video, except for Experiment 4, which looked at how advance information about the task affected responses. Five out of seven experiments manipulated sentence structure, with canonical SVO structure for half the critical items and OVS structure for the other half (see Examples (3) and (4), respectively):

(3) *Der_{Nom} Kellner beglückwünscht den_{Acc} Millionär außerhalb des Geschäfts* (SVO: 'the waiter is congratulating the millionaire outside the shop').

(4) *Den_{Acc} Kellner beglückwünscht der_{Nom} Saxofonist außerhalb des Geschäfts* (OVS, roughly: 'the waiter is being congratulated by the saxophone player outside the shop').

Finally, we manipulated the type of cue that participants received regarding the continuation of the sentence: In all experiments, the speaker and her gaze shift were visible on half the trials and obscured on the other half (Fig. 1). The speaker in Experiments 1-6 was human, and her gaze behaviour was based on typical gaze in speech production (e.g., Griffin & Bock, 2000). The speaker in Experiment 7, by contrast, was a virtual agent, Billie (the BIelefeld Life-Like Interactive agEnt, Fig. 2), whose head and mouth movements were modelled on the timing of our human speaker.

Fig. 2: Example from a video with the virtual agent Billie.

Either in addition to or instead of the speaker's gaze, Experiments 5 and 6 also presented a different type of cue (Kreysa, Nunnemann, & Knoeferle, in progress): One or two objects representing the action described by the verb appeared between the agent and patient characters (the millionaire and the waiter in (3)), thus revealing the referent of the second noun phrase based on verb information (see Knoeferle & Crocker, 2007). Table 1 presents an overview of the factors manipulated across experiments. To date (August 2012), Experiments 1-5 have been completed, while Experiments 6 and 7 are in preparation.

Table 1: Overview of conditions across experiments.

Experiment	Manipulated Factor Levels	Verification Task
1	Gaze *vs.* NoGaze, SVO *vs.* OVS	referents
2		patient
3		role relations
4		referents *vs.* role relations
5	Gaze *vs.* NoGaze, one action *vs.*no action	role relations
6	Gaze *vs.* NoGaze, two actions *vs.* no action	role relations
7	Virtual Agent Gaze *vs.* NoGaze, SVO *vs.* OVS	role relations

In the following, we will present the Methods simultaneously for all experiments, emphasising only important differences. Similarly, we will generalise across experiments in presenting the Results while highlighting conclusions about the effects of the different manipulated conditions.

2. METHODS

2.1 Participants

For each of the seven experiments, we recorded response times and eyetracking data from 32 native German speakers of the Bielefeld University student population (ages 17-35; Mean age to date = 23.5, 30% male). Each participant took part in only one experiment. All reported normal or corrected-to-normal vision and hearing, as well as not speaking a language other than German before age six. They received 6 € or course credit for participation. We have excluded twelve participants out of the 172 tested so far (7%) for technical reasons or low accuracy (< 83%). Overall accuracy in the verification task was high (96-99%).

2.2 Materials

The same 24 experimental items (i.e., screen display + sentence) and 48 filler trials were used for all seven experiments. In experimental trials, the transitive action always takes place between the central character (e.g., the waiter) and one of the two outer characters (the millionaire or the saxophone player). All characters and action objects were created in the virtual world Second Life®, and pre-tested to ensure that they were easily recognisable. All experimental sentences had masculine noun phrases that were unambiguously case-marked for grammatical function. Shortly after pronouncing the verb, the speaker shifted gaze from the central, pre-verbal referent to the post-verbal referent. For each item, if the SVO sentence structure mentioned the post-verbal referent on the right side (e.g., the millionaire), the OVS post-verbal referent was the character on the left (the saxophone player) and vice versa, so that the speaker shifted gaze to the right and left equally often for both sentence structures within a list. In Experiments 5 and 6, the action objects (e.g., a bunch of balloons to represent *congratulate*), appeared between the two characters referenced in the sentence (see Fig. 3; Kreysa, Nunnemann, & Knoeferle, in progress).

Fig. 3: Experiment 5: Example of a trial showing both the speaker and the action-related object.

The same human speaker (PK) recorded all videos, from which the audio files of the sentences were extracted. We hand-coded the onsets of all sentence parts and gaze shifts using ELAN (http://tla.mpi.nl/tools/tla-tools/elan/), and used these cue-points to script the gaze behaviour, speech timing, and speech speed of the virtual agent. Billie uses the Articulated Communicator Engine (Kopp & Wachsmuth, 2004) together with Mary TTS for speech synthesis (http://mary.dfki.de). However, Billie's speech production only served the purpose of generating appropriately timed lip movements: In a second step, we superimposed the human speaker's audio file on the virtual agent videos, to maximise comparability of the human and virtual agent gaze shifts. As a result, the virtual agent's behaviour closely resembled the human speaker's production of our experimental sentences. The onset of the human speaker's gaze shift to the post-verbal referent was also used as the time-point at which the tools in Experiments 5 and 6 appeared.

All experiments used response template screens, which specified the required verification response. The three different types of templates (verification of reference, patient, or role relations) all displayed three stick men on a grey background (Fig. 4). Participants were told that each of these represented the character in that position in the Second Life display during the video. What varied between template types was the blue marking indicating the to-be-verified aspect: For referent verification, two potential referents were circled (Fig. 4a – this would be a "match" response for the scene in Fig. 1 with sentence (3), above); for patient verification, one referent was circled, the to-be-verified patient of the sentence (Fig. 4b – a "mismatch" for sentence (3) but a "match" for (4)). On the role-relations template, an arrow pointed from the agent to the patient (Fig. 4c – this would be a mismatch for both sentences (3) and (4), since the millionaire on the left is never mentioned in agent role). The templates were paired with the sentences in such a way that half of the experimental trials required match responses, the other half mismatches. Experiment 4 constituted an exception, because for counterbalancing reasons all experimental trials in this study were matches. To compensate, 75% of fillers were mismatches, resulting in equal proportions of matches and mismatches across all trials.

Fig. 4: Response templates for verifying (a) the mentioned referents (Experiments 1 and 4), (b) the patient of the sentence (Experiment 2), or (c) the role relations (Experiments 3, 5-7).

2.3 Apparatus and Procedure

Participants' verification responses and reaction times were recorded using a Cedrus® RB series response pad, and their eye movements were tracked with a desktop-mounted

EyeLink® 1000 (SR Research) eyetracker. Once the tracker was set up and calibrated, participants received on-screen instructions and at least four practice trials with feedback in order to acquaint them with the videos and the templates. Each trial began with a central fixation point. For Experiments 1-3 and 5-7, participants next saw the video. It began with the speaker (if visible) smiling into the camera (see Fig. 3) and then looking at each of the three characters in turn. Only then, after about six seconds, did she produce the sentence, looking first at the central character, then shifting gaze to the post-verbal referent. Shortly after the end of the sentence, the video was replaced by the verification template, which remained on-screen until participants pressed the response button. Finally, the fixation point reappeared in preparation for the next trial.

For Experiment 4, the initial fixation point was followed directly by the template (either reference or role relations verification, see Fig. 4(a) and (c), respectively). The template remained on-screen until the participant pressed a button to indicate that s/he had memorised it. This triggered the onset of the video. In this experiment, participants made speeded verification responses: They were asked to press the button as soon as they knew whether the initial template matched the sentence they were hearing, even during the video (it ended as soon as the button was pressed). This contrasts with the other experiments, where the responses always occurred post-video.

After 72 trials, participants took a memory test (not reported here), filled in a debrief questionnaire, and provided basic demographic data. In total, the experiments lasted 45-60 minutes.

2.4 Design

Although the manipulated factors differed, all Experiments were set up with a 2*2*2 design, within participants and items: Combinations of conditions were distributed across eight experimental lists using a Latin-square design, and the order of trials was pseudo-randomised for each participant. Speaker Gaze Availability (Gaze vs. NoGaze) is the key manipulation in all seven experiments (human speaker in Experiments 1-6, virtual agent in Experiment 7). In addition, Experiments 1, 2, 3, and 7 varied Sentence Structure (OVS vs. SVO) and Template Congruence (match vs. mismatch); Experiment 4 varied Sentence Structure and Task (Reference vs. Role Relations). Experiments 5 and 6 compared the effects of two types of contextual cue (Speaker Gaze vs. Depicted Action), so in order not to inflate the number of conditions, Sentence Structure was kept constant (SVO); Template Congruence was varied. Regarding the contextual cues in these two experiments, 25% of trials showed Gaze and Action, 25% of trials NoGaze but Action, 25% of trials showed Gaze but no Action, and finally neither cue appeared on 25% of trials.

Additional comparisons can be made between experiments: The only aspect that differs between Experiments 1-3 is the comprehension task, thus Task can be seen to vary between participants/experiments but within items here, in addition to its within-participant and within-item variation in Experiment 4. Similarly, different types of contextual information are manipulated within participants and items in Experiments 5 and 6 (gaze and/or action/s), but between Experiments 3 (human speaker) and 7 (virtual agent).

2.5 Dependent variables and Analysis

Our primary interest is to show how seeing a speaker's gaze shift or a depicted action affects the attention allocated to the referent of the second noun phrase (NP2), which we will refer to henceforth as the *target character*. If the available contextual cues (gaze or actions) are used to anticipate this referent, this would cause a rise in fixations to the target character, relative to the unmentioned character (the *competitor*). In particular, this might occur before the onset of the NP2. Additionally, a contextual cueing effect could depend on the type and number of available contextual cues (human vs. virtual speaker and/or one or two actions), on the sentence structure (SVO vs. OVS), or on the participant's goal, as determined by the verification task (reference, patient, role relations; after or before comprehension).

With regard to fixation data, we defined two critical time periods: 800 ms between the onset of the speaker's gaze shift and the onset of the NP2 (hereafter: SHIFT period; this is where we might see anticipatory looks) and the first 800 ms of the NP2. Within these two periods, we used linear mixed effects models to analyse the mean log probability ratios for gazes to the target character relative to the competitor, averaged over participants and items. In addition to these log-ratios, we also analysed counts of fixations to the target character in the relevant time periods (hierarchical loglinear analyses), onset times of the first post-shift fixation to the target character, and log-transformed response times. However, in the interest of brevity, the summary of results presents only those that are most pertinent to our conclusions. Full details can be found in Knoeferle and Kreysa (2012).

3. RESULTS & DISCUSSION

3.1 Effects of speaker gaze

All studies revealed a clear effect of speaker gaze on listeners' visual attention: When they could see the speaker's gaze shift, participants tended to fixate the character that would be referred to next even before the onset of its name. Fig. 5 plots proportions of fixations to the target character over time for Experiment 3 (role relations). A similar pattern of fixations was found across experiments.

Clearly, fixations to the target character began to rise steeply as early as 500 ms after the speaker started to shift her gaze. This means that participants frequently began fixating the target character prior to its mention in the sentence, i.e. they anticipated it. By contrast, in the NoGaze conditions the target character was fixated almost a second later, once the sentence disambiguated between the two potential referents.

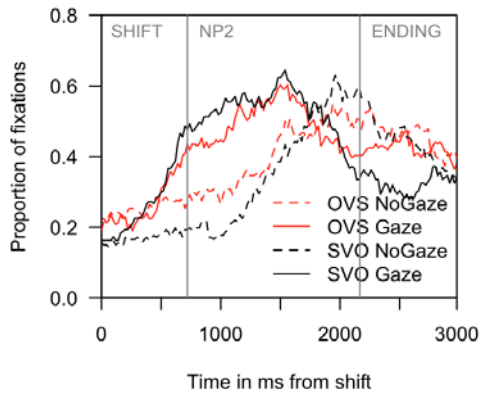

Fig. 5: Experiment 3: Target fixations over time from the onset of the speaker's gaze shift. The grey vertical bars indicate the mean onsets of the NP2 and of the sentence end.

Interestingly, the strong effect of speaker gaze was found despite the fact that the speaker herself was rarely fixated directly during the sentence. In fact, Fig. 6 shows that fixations to the speaker were almost as rare as fixations to the uninformative scene background.

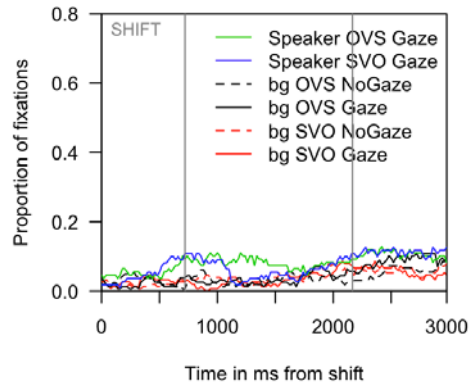

Fig. 6: Experiment 3: Fixations to the speaker and background (bg) over time from the onset of the speaker's gaze shift.

An effect of speaker gaze in the absence of direct speaker looks fits well with findings that gaze-cueing is a low-level effect, which can be found in infants (Scaife & Bruner, 1975; Brooks & Meltzoff, 2002), under cognitive load, or even when presented subliminally (Xu et al., 2011). Our results suggest that peripheral perception of the direction of the speaker's head movement is sufficient to direct attention to where she is looking. This is the case both for a frontal view of the speaker (Hanna & Brennan, 2007; Staudte & Crocker, 2011) and for situations where the speaker is positioned at an angle (see Fig. 1).

Yet although we and others have found a strong effect of speaker gaze in most studies to date, results from Experiment 4 suggest that slight changes in the circumstances (varying comprehension tasks; templates presented prior to the video), can lead to substantially smaller effects. In fact, although the gaze effect does reach statistical significance at least in the NP2 time period ($t > 5.0$), it is hardly visible in the graphs of target fixations, as Fig. 7 reveals.

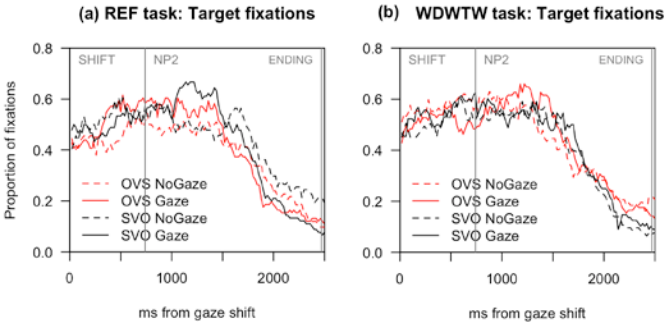

Fig. 7: Experiment 4: Target fixations over time separately for the tasks of verifying (a) reference and (b) role relations.

One explanation for the reduced gaze effects on anticipatory looks to the target character may be that participants in this experiment were already much more likely to be fixating the target at the time of the speaker's gaze shift (~40% in Fig. 7 vs. ~20% in Fig. 5). This could be because they had top-down, task-based expectations of what they were looking for, i.e. which character was relevant for detecting the match or mismatch of template and sentence. In addition, they could already respond during the sentence, so they may have been motivated to anticipate the post-verbal referent earlier than in the other six experiments. To the extent that they were already looking at the target when the speaker shifted her gaze, this cue would have provided confirmatory information only, not an incentive to shift gaze to a new location. Thus, the combination of prior expectations and early anticipation may arguably have reduced the importance of the gaze cue.

It is conceivable that other factors may also affect a listener's visual attention in response to a gaze cue. Examples could be visual highlighting or movement in the actual scene, visibility of the speaker, but also more social aspects like participants' trust in or knowledge about the speaker (see e.g., the incongruent arrays used in Hanna & Brennan, 2007, their Experiment 2). Identifying situations and contexts where speaker gaze plays an important vs. a minor role for situated comprehension is an interesting challenge for further research.

One such factor that might affect the extent of gaze-following is the identity of the speaker as a human vs. as an artificial agent, which we explore in Experiment 7. The agent setup is characterised by several key differences: For one, only the head of our virtual agent shifts, not its eyes. Despite all efforts to ensure comparable timing, this distinguishes the gaze cue produced by the agent from the human gaze cue. However, since the human speaker's gaze shift seems to be processed peripherally, this may not be problematic: It is unlikely that an eye shift without any head movement can be perceived in the periphery (Loomis et al., 2008).

Another issue is that an interlocutor's gaze direction may be salient because it symbolises his/her intentions (Becchio et al., 2008; Meltzoff et al., 2010; Staudte & Crocker, 2011). Although interacting with an adept virtual agent may share some aspects of interacting with a human – and in fact, such natural interaction is an often-stated goal in designing conversational agents – it is possible that participants may

not ascribe full intentionality to an artificial agent recorded in a video with superimposed sound (we plan to assess the extent to which participants ascribe human traits and intentionality to Billie). In this case, and if attributing intentionality is important for speaker gaze effects, this could lead participants to follow the agent's gaze less than they follow the human speaker's gaze (though note that previous research by others, as well as the consistent congruence of gaze and sentence in our study make it unlikely that there would be no effect of gaze at all). Alternatively, finding no substantial difference in how human and agent gaze shifts are processed would weaken the argument that intentionality is critical to gaze effects.

3.2 Effects of action information

Of course, speaker gaze is not the only contextual cue that can cause a rapid shift in listeners' attention during comprehension. For example, depicted actions have been shown to disambiguate between alternative sentence meanings (Knoeferle et al., 2005; Knoeferle & Crocker, 2006). Experiment 5 pitted speaker gaze against depicted actions, with the aim of providing insight into their relative effects (Kreysa, Nunnemann, & Knoeferle, in progress). Fig. 8 shows how the two contextual cues affect fixations to the target character (Fig. 8a) and to the action object and speaker (Fig. 8b).

The fixation proportions in the Gaze-only and NoCue conditions in Fig. 8a are very similar to those found in the previous experiments. By contrast, the curves in the conditions with a depicted action differ starkly: Recall that the object depicting the action appeared at the same time as the speaker initiated her gaze shift. The sudden onset of the object on the screen attracted attention almost immediately in the majority of trials (see Fig. 8b). This means that the object initially distracted attention *away* from the target character (see the dip in the +Object lines in Fig. 8a), whereas the peripherally processed gaze cue consistently caused a rise in target fixations (see the difference between Gaze-only and NoCue in Fig. 8a).

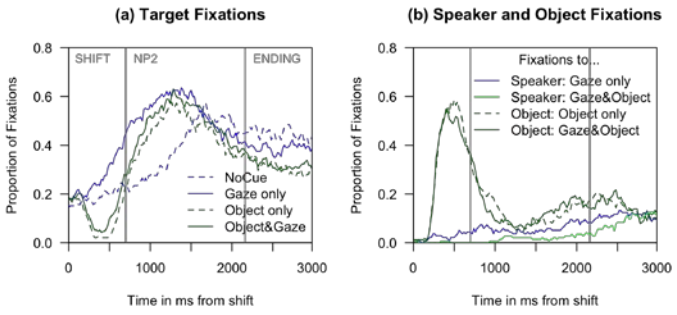

Fig. 8: Experiment 5: Fixations to (a) the target and (b) the speaker and the action object over time

However, Fig. 8a also shows that participants in the +Object conditions rapidly recovered from the distraction and swiftly shifted their gaze to the target character. Already early in the NP2 period, there is no longer any difference between the Gaze and Object cues. By 1000 ms from shift/object onset, both cues seem equally helpful for locating the target, and both elicit more/earlier target fixations than the NoCue condition. Interestingly however, there is no additive effect: It seems that one cue is necessary but also sufficient to achieve disambiguation.

Experiment 6 (in preparation) serves to clarify further how the action cue is used: Presenting an object on both sides of the agent character (a verb-congruent object on the side of the target, an incongruent one on the side of the competitor) will inform us of the extent to which the semantic identity of the object is processed, rather than just the fact that it points towards the target.

3.3 Effects of sentence structure

In Fig. 5, above, the proportions of fixations to the target character differ clearly between sentence structures, with a stronger effect of Gaze for SVO (17% more fixations with Gaze than NoGaze) than OVS sentences (6%). Fixations in Experiment 1 showed a similar pattern, though somewhat weaker (Knoeferle & Kreysa, 2012). This ties in well with the idea that OVS sentences are harder to process, so there may be fewer available cognitive resources for integrating an additional source of information than in the SVO case. However, in both experiments the interaction of gaze and structure was largely due to the unexpected finding that even in the baseline (NoGaze), the target character was fixated earlier for OVS than SVO sentences.

The clearest effect of structure comes from Experiment 2, where participants were preparing to verify the patient (the accusative object) of the sentence. Fig. 9 shows quite different patterns of fixations for the two sentence structures in this task: In OVS sentences, this patient is known from the initial determiner *Den*; and participants in this condition seemed reluctant to shift gaze away from this character.

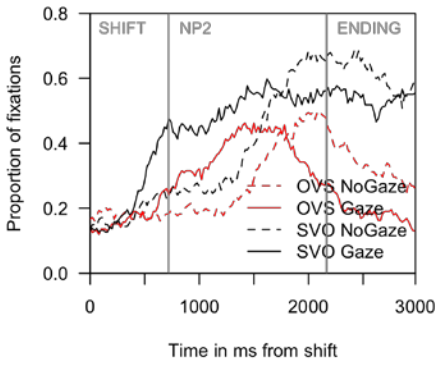

Fig. 9: Experiment 2: Target fixations over time from the onset of the speaker's gaze shift.

Although they did look at the post-verbal referent – the subject and agent of the sentence – they did so later than during SVO sentences and only briefly, as shown in the rapid decrease in fixations at the end of the sentence. By contrast,

in SVO sentences the patient is mentioned post-verbally: Here, participants were quick to shift gaze to this character, and they continued fixating it until the verification template appeared.

3.4 Effects of the verification task

Since the type of verification task was the only difference between Experiments 1-3, the variation in the effect of structure and/or in its interaction with speaker gaze (see previous section) must be due to the task differences. This suggests that online comprehension can be affected by the comprehender's current goal: It is not entirely independent of subsequent tasks (for similar effects of affordances, see Chambers et al., 2004; for effects of mental representations, see Altmann & Kamide, 2009; and for online ERP effects of sentence-verification, see Knoeferle et al., 2011).

Yet visual-world paradigm studies in the previous decades have used a variety of diverse tasks, such as carrying out instructions to move objects (e.g., Allopenna et al., 1998; Spivey et al., 2001) versus listening to understand (e.g., Kamide, Scheepers, & Altmann, 2003; Knoeferle & Crocker, 2006). The roughly consistent patterns of eye movements found across studies led to the implicit assumption that the effects under examination would generalise across tasks (Altmann, 2011; Salverda et al., 2011). Because current accounts of situated comprehension (e.g., Altmann & Kamide, 2007; Knoeferle & Crocker, 2006) and associated computational models (Mayberry et al., 2009) are based on the results of these studies, they also lack any explicit consideration of potential task or goal effects (see Salverda et al., 2011).

In fact, our results show in two different senses that task *does* matter. The first has already been mentioned: Comparing Experiments 2 and 3, the change in the post-sentence task between verifying either just the identity of the patient or the full thematic role structure led to fundamentally different patterns of fixation, both when anticipating the postverbal referent (SHIFT period) and at the end of the sentence. We argue that the two distinct fixation patterns index differences in the underlying comprehension processes. Since the patient verification task of Experiment 2 is (arguably) one sub-component of verifying thematic role relations (Experiment 3), it is interesting to speculate on whether the difference between the two patterns might point to the "missing" subtask of verifying the agent. To date, we have not run such an experiment, but the idea of isolating comprehension sub-tasks in this way seems a promising avenue for further research.

Another effect of task emerges when considering Experiment 4: Here, task was explicitly varied within the experiment. We found more target fixations for the role-relations task than for reference verification ($t = 2.3$), but the pattern of fixations did not differ fundamentally between the two tasks in this experiment (see Fig. 7). Comparing between experiments however, the difference in fixation patterns is striking (cf. Fig. 5; Experiment 3). The videos in the two experiments were identical, but the mere change of presenting the response template before vs. after the video fundamentally altered the pattern of attention: Experiment 4, where participants had memorised the response template and responded quickly, found much greater anticipation of the target character across conditions, compared to Experiment 3, where anticipation was limited to trials in which the speaker was visible.

4. CONCLUSIONS

In summary, we show that online comprehension of identical linguistic material is affected by multiple interacting aspects of the context in which it is processed. One such aspect of the context is the goal of the specific instance of comprehension. In empirical studies of language comprehension, this goal will often be defined by the experimental task. Another cue is speaker gaze, which affects visual attention rapidly and peripherally. If this specialised processing of the direction of attention extends to gaze from artificial agents, speech-related gaze behaviour would be a highly desirable feature for designing communicative agents with which humans can interact in a natural and resource-efficient way. We show that speaker gaze effects are robust across two sentence structures and a variety of verification tasks, but also that they are in turn modulated by structure and task.

It makes sense to compare the different kinds of contextual cues with regard to the extent and time-course with which they direct listeners' attention to upcoming sentence content, and the situations in which this occurs. Experiment 5 showed that although verb-related information in visual context is not necessarily processed in the same way as speaker gaze, it can have the same ultimate effect of allowing anticipation. In terms of the CIA (Knoeferle & Crocker, 2006; 2007), the facilitatory effect of action information on target fixations can be conceived of as utterance-mediated attention: The action object appears while participants are processing the verb it relates to. This allows them to update their model of the scene, integrate verb and action, and to rapidly use this cue to anticipate the post-verbal referent.

In fact, current processing accounts of situated comprehension (Altmann & Kamide, 2007; Knoeferle & Crocker, 2006; 2007) already include verb-mediated attention. In contrast, speaker gaze has not yet been accommodated. Precisely because its effect is rapid and robust, it would seem important to integrate speaker gaze into these accounts. One possibility is that the gaze effect bypasses the steps of utterance-mediated attention and language-scene integration: The speaker's gaze shift could pull the listener's attention to the target character directly, without being mediated by the lexical term or integrated with the unfolding interpretation. Alternatively, a speaker's gaze shifts might be integrated more closely with the unfolding sentence interpretation.

This is just one example of how the robust effects of speaker gaze might be interpreted theoretically, within current accounts of situated comprehension. In addition, the accounts will need to specify in more detail which aspects of visual context effects are robust across tasks and linguistic structure, and which ones vary between them. Extending our

knowledge of the time course and situational constraints of speaker gaze use in humans in this way will also allow us to determine key aspects for implementing naturalistic referential gaze behaviour in artificial agents. Overall, our experiments highlight the importance of such extensions by showing just how context-dependent language processing really is.

ACKNOWLEDGMENTS

This research was funded by the German research foundation (DFG) through the Cognitive Interaction Technology Excellence Center (CITEC) at Bielefeld University. We thank Eva Mende, Anne Kaestner, Eva Nunnemann, Lydia Diegmann, Kristin Kleinehagenbrock, Katja Münster, and especially Linda Krull for assistance with preparing the stimulus materials and/or collecting data.

REFERENCES

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

Altmann, G.T.M. (2004). Language-mediated eye movements in the absence of a visual world: The 'blank screen paradigm'. *Cognition, 93,* B79-B87.

Altmann, G.T.M. (2011). The mediation of eye movements by spoken language. In S.P. Liversedge, I.D. Gilchrist, & S. Everling (Eds.), *The Oxford Handbook of Eye Movements* (pp. 979-1003). Oxford, UK: OUP.

Altmann, G.T.M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57,* 502-518.

Altmann, G.T.M., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition, 111,* 55-71.

Bailenson, J., & Yee, N. (2005). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science,* 16, 814-819.

Baldwin, D.A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language, 20,* 395-418.

Baldwin, D.A. (1995). Understanding the link between joint attention and language. In C. Moore and P.J. Dunham (Eds.), *Joint attention. Its origins and role in development* (pp. 131-159). Hillsdale, NJ: LEA.

Bavelas, J., & Chovil, N. (2000). Visible acts of meaning. *Journal of Language and Social Psychology, 19,* 163-194.

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

Bergmann, K., Kopp, S., & Eyssel, F. (2010). Individualized gesturing outperforms average gesturing: Evaluating gesture production in virtual humans. *IVA 2010, Lecture Notes in Computer Science, 6356,* 104-117.

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

Brooks, R., & Meltzoff, A.N. (2002). The importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology, 38,* 958-966.

Castelhano, M.S., Wieth, M., & Henderson, J.M. (2007). I see what you see: Eye movements in real-world scenes are affected by perceived direction of gaze. In L. Paletta and E. Rome (Eds.), *WAPCV 2007, Lecture Notes in Artificial Intelligence 4840* (pp. 251-262). Berlin: Springer.

Chambers, C.G., Tanenhaus, M.K., & Magnuson, J.S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30,* 687-696.

Clark, H.H. (1996). *Using language*. Cambridge: Cambridge University Press.

Crocker, M.W., Knoeferle, P., & Mayberry, M.R. (2009). Situated sentence processing: The coordinated interplay account and a neurobehavioral model. *Brain and Language, 112,* 189-201.

Deubel, H., & Schneider, W.X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research,* 36, 1827-1837.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition, 6,* 509-540.

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

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

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

Kamide, Y., Altmann, G.T.M., & Haywood, S.L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49,* 133-156.

Kamide, Y., Scheepers, C., & Altmann, G.T.M. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research, 32,* 37-55.

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

Kleinke, C.L. (1986). Gaze and eye contact: A research review. *Psychological Bulletin, 100,* 78-100.

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

Knoeferle, P., & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: Evidence

from eye movements. *Journal of Memory and Language, 57,* 519-543.

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

Knoeferle, P., Habets, B., Crocker, M.W., & Münte, T.F. (2008). Visual scenes trigger immediate syntactic reanalysis: Evidence from ERPs during situated spoken comprehension. *Cerebral Cortex, 18,* 789-795.

Knoeferle, P., & Kreysa, H. (2012). Can speaker gaze modulate syntactic structuring and thematic role assignment during spoken sentence comprehension? *Frontiers in Psychology, 3,* 538. doi: 10.3389/fpsyg.2012.00538.

Knoeferle, P., Urbach, T.P., & Kutas, M. (2011). Comprehending how visual context influences incremental sentence processing: Insights from ERPs and picture-sentence verification. *Psychophysiology, 48,* 495-506.

Kopp, S. (2010). Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. *Speech Communication, 52,* 587 - 597.

Kopp, S., & Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds, 15,* 39–52.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLoS ONE, 3,* e2597.

Kreysa, H., Knoeferle, P., Yaghoubzadeh, R., & Kopp, S. (in progress). Speaker gaze effects and sentence structures: Eye-tracking evidence from an experiment with a virtual speaker.

Kreysa, H., Nunnemann, E.M., & Knoeferle, P. (in progress). Visual context cues for spoken sentence comprehension: Speaker gaze and depicted actions.

Loomis, J.M., Kelly, J.W., Pusch, M., Bailenson, J.M., & Beall, A.C. (2008). Psychophysics of perceiving eye-gaze and head direction with peripheral vision: Implications for the dynamics of eye-gaze behavior. *Perception, 37,* 1443-1457.

Mayberry, M.R., M.W. Crocker and P. Knoeferle (2009). Learning to attend: A connectionist model of situated language comprehension. *Cognitive Science, 33,* 449-496.

Meltzoff, A.N., Brooks, R., Shon, A.P., & Rao, R.P.N. (2010). "Social" robots are psychological agents for infants: A test of gaze following. *Neural Networks, 23,* 966-972.

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

Ricciardelli, P., Bricolo, E., Aglioti, S., & Chelazzi, L. (2002). My eyes want to look where your eyes are looking: Exploring the tendency to imitate another individual's gaze. *NeuroReport, 13,* 2259-2264.

Richardson, D.C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29,* 1045-1060.

Richardson, D.C., Dale, R., & Kirkham, N.Z. (2007). The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological Science, 18,* 407-413.

Salverda, A.P., Brown, M., & Tanenhaus, M.K. (2011). A goal-based perspective on eye movements in visual world studies. *Acta Psychologica, 137,* 172-180.

Scaife, M., & Bruner, J.S. (1975). The capacity for joint visual attention in the infant. *Nature, 253,* 265-266.

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

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

Spivey, M.J., Tyler, M.J., Eberhard, K.M., & Tanenhaus, M.K. (2001). Linguistically mediated visual search. *Psychological Science, 12,* 282-286.

Staudte, M., & Crocker, M.W. (2011). Investigating joint attention mechanisms through spoken human-robot interaction. *Cognition, 120,* 268-291.

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

Weber, A., Grice, M., & Crocker, M.W. (2006). The role of prosody in the interpretation of structural ambiguities: a study of anticipatory eye movements. *Cognition, 99,* B63-B72.

Xu, S., Zhang, S., & Geng, H. (2011): Gaze-induced joint attention persists under high perceptual load and does not depend on awareness. *Vision Research, 51,* 2048-2056.

Yu, C., Ballard, D.H., & Aslin, R.N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science, 29,* 961-1005.