frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 18 April 2023
DOI 10.3389/fcomm.2023.1029157



OPEN ACCESS

EDITED BY
Juhani Järvikivi,
University of Alberta, Canada

REVIEWED BY
James P. Trujillo,
Radboud University, Netherlands
Lucas Battich,
École Normale Supérieure, France

*CORRESPONDENCE
Eva M. Nunnemann
✉ enunnemann@gmail.com
Pia Knoeferle
✉ pia.knoeferle@hu-berlin.de

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

RECEIVED 26 August 2022
ACCEPTED 13 March 2023
PUBLISHED 18 April 2023

CITATION
Nunnemann EM, Kreysa H and Knoeferle P (2023) The effects of referential gaze in spoken language comprehension: Human speaker vs. virtual agent listener gaze.
*Front. Commun.* 8:1029157.
doi: 10.3389/fcomm.2023.1029157



# The effects of referential gaze in spoken language comprehension: Human speaker vs. virtual agent listener gaze

Eva M. Nunnemann[1]*, Helene Kreysa[2] and Pia Knoeferle[3,4,5]*

[1]Independent Researcher, Bielefeld, Germany, [2]Institute of Psychology, Friedrich Schiller University Jena, Jena, Germany, [3]Department of German Studies and Linguistics, Humboldt-Universität zu Berlin, Berlin, Germany, [4]Berlin School of Mind and Brain, Berlin, Germany, [5]Einstein Center for Neurosciences Berlin, Berlin, Germany

**Introduction:** Four studies addressed effects of human speaker gaze vs. virtual agent listener gaze on eye movements during spoken sentence comprehension.

**Method:** Participants saw videos in which a static scene depicting three characters was presented on a screen. Eye movements were recorded as participants listened to German subject-verb-object (SVO) sentences describing an interaction between two of these characters. Participants' task was to verify whether the sentence matched a schematic depiction of the event. Two critical factors were manipulated across all four experiments: (1) whether the human speaker—uttering the sentence—was visible, and (2) whether the agent listener was present. Moreover, in Experiments 2 and 4, the target second noun phrase (NP2) was made inaudible, and in Experiments 3 and 4, the gaze time course of the agent listener was altered: it looked at the NP2 referent about 400 ms before the speaker did. These manipulations served to increase the value of the speaker's and listener's gaze cues for correctly anticipating the NP2 referent.

**Results:** Human speaker gaze led to increased fixations of the NP2 referent in all experiments, but primarily after the onset of its mention. Only in Experiment 3 did participants reliably anticipate the NP2 referent, in this case making use of both the human speaker's and the virtual agent listener's gaze. In all other cases, virtual agent listener gaze had no effect on visual anticipation of the NP2 referent, even when it was the exclusive cue.

**Discussion:** Such information on the use of gaze cues can refine theoretical models of situated language processing and help to develop virtual agents that act as competent communication partners in conversations with human interlocutors.

KEYWORDS

**spoken language, comprehension, referential gaze, virtual agent, speaker gaze, human-agent interaction, eye tracking, eye gaze**

## 1. Introduction

"Gaze is an extremely powerful expressive signal that is used for many purposes, from expressing emotions to regulating interaction" (Lance and Marsella, 2010, p. 50). These regulating functions of gaze comprise, for example, the organization of turn taking, the search for feedback, and the means to emphasize parts of an utterance (Lee and Marsella, 2006). Moreover, in face-to-face interaction, gaze is an important signal for detecting an

interlocutor's focus of attention (e.g., Argyle and Cook, 1976; Fischer and Breitmeyer, 1987; Steptoe et al., 2009). Often gaze also plays a crucial role in the comprehension of spoken language. For example, in joint-search tasks, gaze has proven to be helpful when participants collaborated in finding a specific object (Brennan et al., 2008). Brennan et al. (2008) have shown that gaze was highly efficient for the mediation of collaboration in a spatial task. That listeners are also able to quickly make use of a speaker's gaze for language comprehension was observed by Hanna and Brennan (2007) in two experiments in which speaker gaze disambiguated a target object before it was mentioned in a sentence. Furthermore, speaker gaze can have a facilitating effect on the understanding of event roles (Kreysa and Knoeferle, 2011b).

Not only are people able to detect and make use of gaze cues in human–human interaction, but also in human–robot interaction. Staudte and Crocker (2011) found that people were able to establish joint-attention with a robot as well. Even though head and eye movements of the robotic agent in their experiments were rather rudimentary, participants could make out the object at which the robot gazed. Boucher et al. (2012) showed in one of their experiments on human–robot collaboration that people even learned to infer the direction of a robot's gaze from its head movement alone in a condition in which the robot's eyes were hidden behind sunglasses.

With regard to artificial gaze, it has been shown that people also react to the gaze of a virtual agent (e.g., Raidt et al., 2005; Andrist et al., 2012). Generally, studies from this field aim at the development of human-like gaze behavior for virtual agents by investigating the impact of agent gaze on humans (Bee and André, 2008). However, Andrist et al. (2012) report that the display of poor or unnatural gaze behavior in an agent can have worse effects on interaction than no implemented gaze behavior at all. Overall, various aspects of gaze behavior in human-agent interaction have been addressed. Among these were, for example, the functions that gaze aversion has in terms of regulating the flow of a conversation (Andrist et al., 2013) or establishing rapport with a listening agent (Heylen et al., 2007; Wang and Gratch, 2010). As virtual agents are mainly used in teaching and learning environments, referential gaze in these situations is also a prominent topic in research (Johnson et al., 2000; Bee and André, 2008; Pfeiffer-Leßmann and Wachsmuth, 2008). Martinez et al. (2010) have shown that a fully animated virtual agent displaying gaze behavior attracts participants attention much faster than agents with either a static gaze or one with stepped gaze behavior (consisting of two images). Overall, the display of human-like gaze behavior in a virtual agent proved to be helpful for communication, often facilitating task performance or enhancing learning (Maatman et al., 2005; Raidt et al., 2005). Moreover, agents that show gaze behavior are perceived as more autonomous and natural than agents without gaze behavior (Maatman et al., 2005; Courgeon et al., 2014).

As intelligent virtual agents are used in more and more fields of everyday life, and thus to meet the requirements for different and often specific tasks, research on embodied virtual agents offers a great variety of possibilities for further investigation. Currently, agents are mainly employed in the realm of education and learning as teachers, tutors, or trainers. For these tasks, aspects in the agents appearance and behavior that might help learners or users to comprehend, memorize, and recall the content are crucial. The role agent gaze plays in these processes has been investigated by Andrist et al. (2012). Their results revealed that people recalled the learning content better when the virtual agent gazed at the learning materials (in this case a map) while giving a lesson on the history of China than when the agent exclusively gazed at the participant. An aspect that Andrist et al. (2012) did not manipulate in detail was where exactly the agent looked on the map, i.e., the agent gazed toward the map but did not fixate specific points (e.g., a city) while speaking. Furthermore, like other studies that look at aspects and effects of virtual agent gaze (e.g., Bee and André, 2008; Martinez et al., 2010), Andrist et al. (2012) did not deal with the question of whether agent gaze—being generally used as a cue—is used in the same way as human gaze.

These are open questions, which we investigated in four studies, designed as a series of eye tracking experiments that maximally contrasted human and virtual agent gaze when both were present as interlocutors at the same time. That means both gaze cues (human and agent) were available simultaneously. Just as in a natural communication situation with its different communicative roles, the human interlocutor is the speaker while the agent has the role of a listener. With this study approach, it shall be assessed whether people perceive and make use of virtual agent gaze in a similar way as they would exploit human gaze. The series of experiments described in the following investigated the effects of the two different gaze types, i.e., human speaker gaze and virtual agent listener gaze, on spoken language comprehension.

## 2. Experiments

In four eye tracking experiments, participants watched videos in which a human speaker and an agent listener jointly looked at a computer screen displaying a static scene with three clearly distinguishable characters. Simultaneously, the human speaker uttered grammatically correct German SVO sentences, like *Der Kellner beglückwünscht den Millionär* ("The waiter congratulates the millionaire"), describing an interaction between two of the characters from the scene. Meanwhile, the human speaker looked at the referents about 200 *ms* before she mentioned them. The agent listener reacted to her gaze 400 *ms* later and followed it. After each video, a gray template with three stick men representing the characters from the scene and a blue arrow representing the direction of the described interaction appeared on screen. Participants' task was to verify whether it correctly described the sentence/video. After the completion of this first eye-tracking part of the studies, participants solved a gated memory task. Here, they were asked to recall the item sentences in three steps. They were shown a picture of the noun phrase 1 (NP1) referent (e.g., "the millionaire") and had to name the verb as well as the noun phrase 2 (NP2). In case they had trouble recalling the NP2 target character, participants were shown a hint in the form of three possible NP2 targets. After that, participants rated the virtual agent according to warmth and competence. In all four eye tracking experiments, we manipulated three factors: whether the speaker was visible, whether the agent listener was visible, and whether the template matched the video clip. Experiment 1 constituted the basic study.

In Experiments 2 and 4, the NP2 was overlaid with pink noise and was thus inaudible. The purpose of masking the NP2 with noise was to boost participants' potential reliance on the virtual agent listener and human speaker gaze cues for their anticipation of the millionaire. Moreover, in Experiments 3 and 4, the gaze time course of the agent listener was altered, so that it looked at the referents 400 *ms* before the human speaker. This latter change had the purpose of boosting the participants' reliance on the agent listener gaze.

## 2.1. Method and design

### 2.1.1. Participants

A total of 128 monolingual German native speakers aged between 18 and 30 years (mean age 22.9; *N* female = 91) took part in Experiments 1–4 which means that 32 participants took part in each experiment.[1] Each participant could only participate in one of the four experiments. Four participants had to be replaced: two for being bilingual, one for lying about age, and one for technical issues. Their sight was normal or corrected-to-normal. For their participation, they were either paid €6 or they were credited one test person hour. They provided written consent to participate in the study and for the collected data to be used for scientific purposes. The experiments were approved beforehand by the ethics commission at Bielefeld University.

### 2.1.2. Materials

For the series of four experiments, 24 item videos as well as four practice videos were produced. Part of the materials for all these clips came from Kreysa and Knoeferle (2011a), who had created videos displaying a computer screen with three clearly identifiable static characters placed on a landscape and a human speaker sitting to the right of this screen. The characters for their 24 critical items as well as practice trials came from the online platform SecondLife. The remaining characters originate from clipart programs. Screenshots of all characters were pretested by Kreysa and Knoeferle (2011a) to ensure that participants could accurately recognize them.

Each of the 24 critical item videos was accompanied by a grammatically correct, unambiguous German subject-verb-object (SVO) sentence describing a transitive action between the character visible in the middle of the screen (e.g., the waiter) and one of the two outer characters (e.g., the saxophone player and the waiter). An example would be *Der Kellner beglückwünscht den Millionär* ("The waiter congratulates the millionaire."; see Figure 1). These core sentence beginnings were followed by a sentence ending such as *vor dem Geschäft* ("outside the shop"). In Experiments 2 and 4, the noun phrase 2 (NP2, e.g., *den Millionär*) was overlaid by pink noise to make it inaudible.

In each of the video clips, the speaker was positioned next to the screen at an angle that allowed participants to see her face and eye movements throughout the whole clip. She always looked at the camera first—smiling at the participant—before she turned toward the screen inspecting each of the three characters in a fixed order. Subsequently, she turned her gaze again to the central character, e.g., the waiter, which was always the noun phrase 1 (NP1) referent of the subject-verb-object (SVO) sentence that she was about to utter. During the utterance, she always looked at the respective characters in turn, displaying a gaze shift from the NP1 referent toward the NP2 referent shortly after producing the verb. As with the characters, Kreysa and Knoeferle (2011a) had also pretested for each video whether people could detect which character the speaker's gaze was directed at.

For the present experiments, we set out to embed these "speaker videos" into new video clips showing the virtual agent Billie. In the resulting video clips for the first two experiments, the virtual agent was given the role of the listener by following the human speaker in her gaze and head movements. In Experiments 3 and 4, agent listener's gaze preceded the human speaker's gaze. To obtain an exact time course for the materials from Kreysa and Knoeferle (2011a), the speaker's nonverbal behavior was transcribed in ELAN (Wittenburg et al., 2006). This procedure allowed us to extract an exact time course for each item's corresponding speaker gaze. With the data from the transcription, we then calculated an appropriate time course for agent listener gaze behavior relative to speaker gaze toward the characters mentioned in the spoken sentences. By time course of the "listener gaze," we mean the relative delay with which the agent followed the speaker's gaze in Experiments 1 and 2 and preceded it in Experiments 3 and 4.

Thus, we reproduced the speaker's gaze behavior for the virtual agent Billie—with the only difference that its gaze and smiles were delayed by 400 *ms* (in Experiments 1 and 2) or preceded (in Experiments 3 and 4). The time course of both gaze types for Experiments 1 and 2 can be seen in Figure 2A. The altered time course for Experiments 3 and 4 is displayed in Figure 2B. The virtual agent was recorded separately and the videos displaying the speaker and the screen were then integrated into the agent videos. We beveled the embedded video at an angle of 40° to make it possible for Billie to gaze at the speaker as well as the characters depicted in the videos (from Kreysa and Knoeferle, 2011a). This setup also allowed people watching the clips to see the interlocutors' faces throughout the video to make out gaze shifts. All materials were pretested to make sure that gaze movement was clearly identifiable (see Nunnemann, 2021, for further details).

### 2.1.3. Design

The design of the four experiments comprised three within-subject factors—two of which came from the video clips (see Figure 3). The first one is *Speaker Gaze* with the two levels, *speaker gaze* and *no speaker gaze*. The second factor is *Agent Gaze*, and corresponding to *Speaker Gaze*, it has two levels *agent gaze* and *no agent gaze*. A third factor is *Congruency* between the content of the spoken sentence and a response template after each video from the

1 Sample size was based on previous studies (Kreysa and Knoeferle, 2011a; Knoeferle and Kreysa, 2012) and it was kept constant across all four experiments to make results comparable. The numbers of participants break down as follows: For Experiment 1, the mean age was 23.0 (*N* female= 24); for Experiment 2, it was 23.0 (*N* female= 20); for Experiment 3, it was 22.8 (*N* female= 27); and for Experiment 4, the mean age was 22.8 years (*N* female = 20).

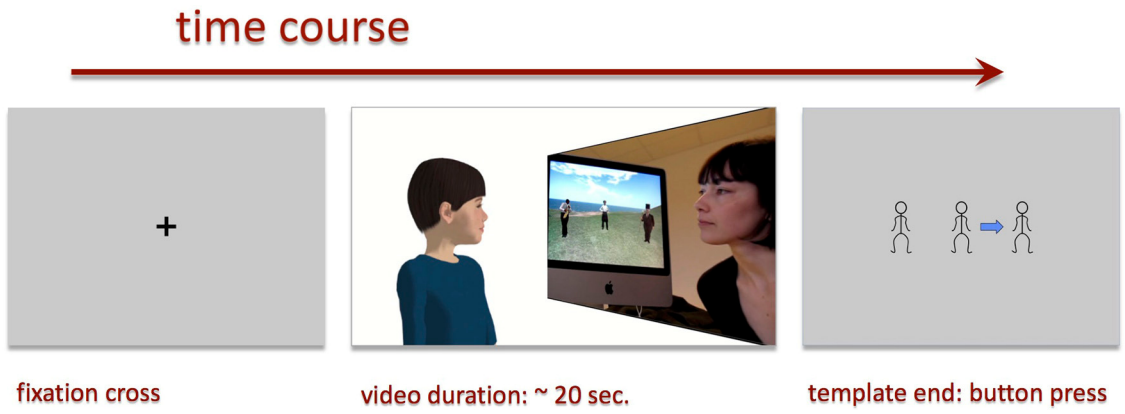


FIGURE 1
The time course of a typical trial in all four experiments: the fixation cross, the video clip, and the template. Adapted from Nunnemann (2021, p. 71).

verification task that participants answered but that is not further discussed in this article.[2]

The two Gaze conditions were distributed over the experiment in such a way that in 50% of all videos, the human speaker was visible, while in the other 50%, she was obscured. Similarly, the virtual agent listener was only visible in half of all videos (see Figure 3). The overall configuration of visibility was distributed in such a way that 25% of clips showed no interlocutor, while in another 25%, both were visible. In addition, the referent of the NP2 appeared equally often to the right and to the left of the NP1 referent in the middle. This means that the human speaker and the virtual agent shifted their gazes equally often to the left and to the right. Randomization of the final lists was done using a *Latin Square* design (see Richardson, 2018).

### 2.1.4. Procedure

Each of the four experiments consisted of two parts. In the first part, participants watched the item videos and solved the verification task after each trial. In the second part, participants performed a gating memory task. An Eyelink 1,000 desktop head-stabilized tracker (SR Research) monitored participants' eye movements and recorded the response latencies after each video clip during the first part of the experiment (movements of the right eye were recorded). The stimuli were shown on a computer screen with a resolution of $1,680 * 1,050$ pixels.

After having given written consent, participants were instructed in all four experiments to watch the videos closely and try to understand videos and sentences as best as they could. Participants were further informed about the template verification task after each trial as well as the post-experiment gated memory task. Then, we explained the verification task to them with an example. Next, we calibrated participants for the eye tracking part of the experiments and they had to complete four practice items. When they had understood the task, the experimenter performed another calibration and started the experiment.

Each of the 24 trials followed the same structure (see Figure 1). Before each trial, a fixation cross appeared in the center of the screen, on which participants were instructed to fixate. Participants watched a video in which the static screen with the three characters and nobody, one of the two interlocutors (human speaker or agent listener), or both interlocutors were visible, and heard a sentence describing an interaction between two of the characters in the scene. After each of these videos, a gray template appeared on the screen depicting the static scene from the video schematically. Three stick men represented the three Second Life characters from the previously seen video and a blue arrow indicated the directionality of the action between the two of them. Participants' task was to decide *via* button press on a CEDRUS box whether the blue arrow represented the action correctly. This so-called verification task (see also Carpenter and Just 1975) was adapted from Kreysa and Knoeferle (2011a).

After participants had completed the eye-tracking part, they completed the gated memory test. As the following analyzes will exclusively focus on the eye tracking results, please see Nunnemann (2021) for details on the further parts of the experiments and the questionnaires which followed. After being either paid or credited for their participantion hour, they were debriefed.

### 2.1.5. Expectations

As the two different gaze types—i.e., human speaker and virtual agent listener gaze—are the major factors, the expectations arising from the design and procedure of the four experiments will be discussed with respect to the different gaze conditions. For the baseline in which neither human speaker gaze or agent listener gaze were present (see Figure 3), participants could solely rely on the spoken sentence to make out the NP2 referent. For Experiments 1 and 3, this means that participants were not able to anticipate the NP2 referent but could only be expected to start fixating it after the NP2 onset. In the two experiments in which the NP2 was overlaid with pink noise (Experiments 2 and 4), we expected them to look equally often at the NP2 target and the competitor.

In the condition in which human speaker gaze was exclusively visible, we expected that participants would make use of speaker

2 See Nunnemann (2021) for further details. *Congruency* plays a role for the verification task that participants answered after each video.

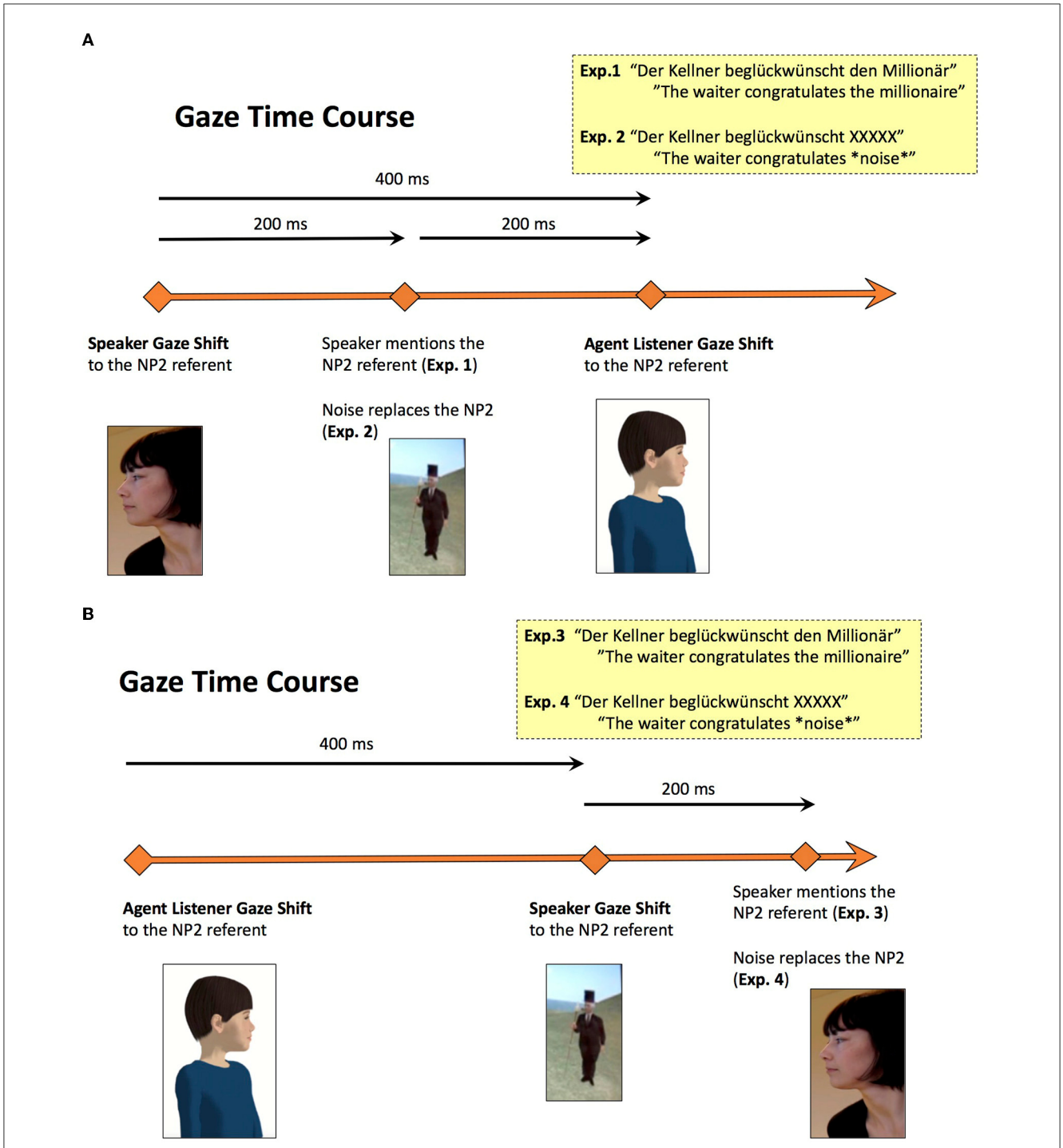


FIGURE 2
Display of the human speaker and agent listener gaze time course. Speaker Gaze Shift and Agent Listener Gaze Shift mark the points when the interlocutors first fixated the NP2 referent **(A)** Experiments 1 and 2. Adapted from Nunnemann (2021, p. 68). **(B)** Experiments 3 and 4. Adapted from Nunnemann (2021, p. 124).

gaze and thus anticipate the NP2 target character before it was mentioned (Experiments 1 and 3) or before the pink noise began, rendering the NP2 inaudible (Experiments 2 and 4). In the latter case, speaker gaze constituted the only cue for the target referent.

In those cases in which virtual agent listener gaze was the only visual cue, we expected for Experiments 1 and 2—where the agent's gaze was delayed about 400 *ms*—that participants would not anticipate the NP2 target referent. Here, we expected that if people reacted to the agent's gaze, this might be due to its novelty effect (see Rehm and André, 2005). For Experiments 3 and 4—in which the agent listener's gaze preceded the human speaker's gaze by 400 *ms*—the expectation was that if participants exploited the agent's gaze cues they would anticipate the target referent. In Experiments 2 and 4—where the critical NP2 was overlaid by pink noise—participants

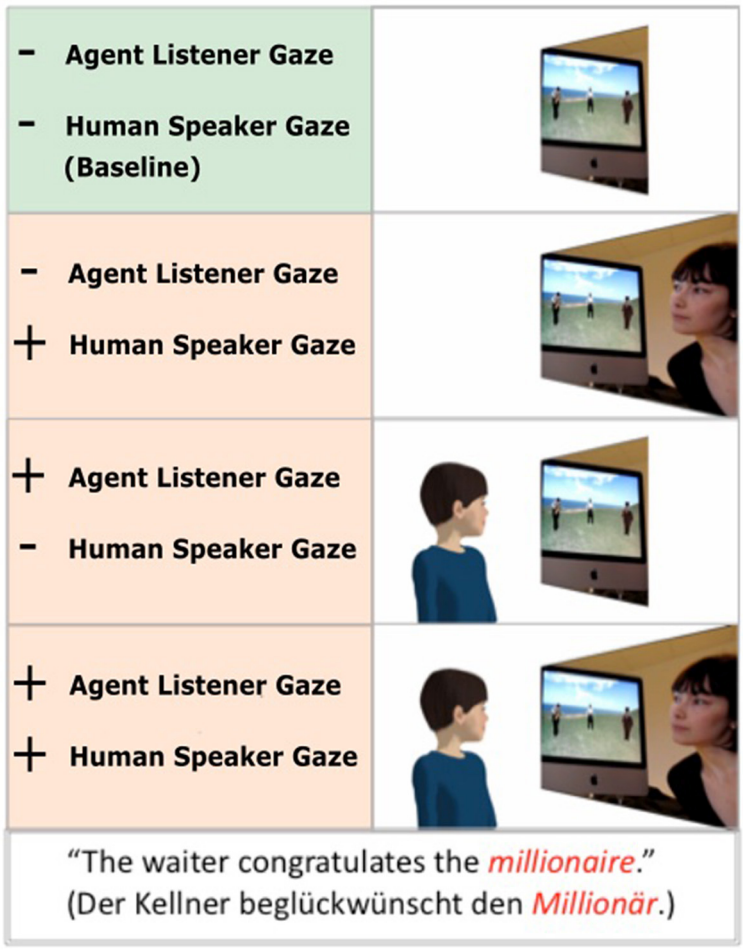


FIGURE 3
Overview of the four conditions used in Experiments 1–4. Adapted from Nunnemann (2021, p. 69).

should exploit the agent's gaze as this was the only cue toward the NP2 referent.

In the condition in which both gaze cues—human speaker and virtual agent listener—were co-present, a variety of outcomes seemed plausible. First, we might observe that the simultaneous presence of both gazes would boost the looks to the NP2 target referent. Alternatively, we might also find that the exploitation of human speaker gaze becomes more difficult when both interlocutors' gazes were visible. Another alternative finding in this condition could be that participants ignored one gaze type. Most likely this would be agent gaze because it had the role of a passive listener, and in Experiments 1 and 2, the agent's gaze followed the speaker's. In Experiments 3 and 4, the agent's gaze preceded the human's, and thus provided new information. Here, it was possible that speaker gaze would be ignored.

## 2.2. Analysis

To analyze participants' eye movement data and to determine which character or interlocutor on the screen was fixated, we created five rectangular areas of interest (AOIs). These regions on the screen were defined as rectangular shapes around the human speaker, the virtual agent, and the three characters in the visual scene (the NP1 referent, the NP2 target referent, and the unmentioned competitor). The three smaller AOIs, which were of the same size, were around the three characters. Two bigger AOIs were placed around the human speaker as well as the virtual agent listener. In this way, fixations to each of these AOIs could be counted. Two of these three characters depicted in the static scene were mentioned in the spoken sentence, which described an interaction between the two. However, the action itself was not depicted in any way. The referent of the NP1 was always the character displayed in the middle. Thus, one of the two outer characters was the NP2 referent. Which of the two characters—to the left or to the right of the NP1 referent—was the NP2 character that remained ambiguous until the human speaker shifted her gaze toward it or referred to it. The third character was the unmentioned distractor. We were mainly interested in participants' gaze behavior in two critical time windows, namely while the speaker shifted her gaze to the target character and when she started to mention the NP2.

The *shift time window* contained all fixations between the beginning of the human speaker's gaze shift toward the NP2 referent and the onset of the NP2 utterance (approx. 719 *ms* after

the shift onset). The *NP2 time window* comprised all the fixations which started in the first 700 *ms* after the NP2 onset, which was uttered by the human speaker in Experiments 1 and 3 and overlaid by pink noise in Experiments 2 and 4. For Experiments 3 and 4, we additionally determined the *agent shift time window* that comprised all fixations between the earlier agent shift onset (the agent shifted its gaze around 400 *ms* before the speaker) and the NP2 onset. These large time windows were then further subdivided into 100 *ms* time bins (named "tiwi" for short in the analyzes).

For the analysis, we did not need to exclude any trials from any of the four experiments due to missing data. Therefore, we inspected the data by looking at descriptive statistics and summaries in *R* (R Core Team, 2017). Furthermore, we did not detect any anomalies that might have caused problems due to participants' blinking behavior.[3] In the analysis, we included all the fixations that fell into the pre-defined time windows. We analyzed the log-gaze probability ratio with which participants were likely to fixate the target character (the referent of the NP2) over the competitor (the unmentioned character, Knoeferle and Kreysa 2012; cf. Kreysa et al. 2018).[4]

For the analysis of these log-gaze probability ratios, we fitted linear mixed effects models for both of the pre-defined time windows, shift and NP2, with random intercepts and slopes for participants and items. As fixed effects, we included speaker gaze and agent gaze, as well as the 100 ms time bins (tiwi) as a third fixed effect, to account for changes in viewing preference over the time window. All three fixed effects were centered. In all analyzes, convergence was achieved for Model 1, the results for which will be reported in the following sections[5]:

$$\begin{aligned} &lograt\_comp \sim cspeaker * cagent * ctiwi + \\ &(1 + cspeaker * cagent|subject) + \\ &(1 + cspeaker * cagent|item) \end{aligned} \quad (1)$$

## 2.3. Results

### 2.3.1. Experiment 1

The aim of Experiment 1 was to examine the effect that two non-linguistic gaze cues—i.e., human speaker gaze and virtual agent listener gaze—have on participants' spoken language comprehension. We accordingly examined participants' visual attention to different pre-defined regions on the screen over time and compared the results for the different conditions.

Figure 4A shows the time course graph for Experiment 1 in successive 20 *ms* time slots. It displays the gaze probability log ratios for the fixations to the target over the competitor character for the entire sentence. From Figure 4A we can see that until the onset of the *shift time window* (from the speaker gaze shift onset until the utterance of the NP2), participants did not look much at either the target (the NP2 referent) or the competitor (the third character). The *log* gaze probability ratios oscillate between -0.5 and 0.5. Only after the onset of the NP2, we can see a clear fixation preference for the target character. For those conditions in which speaker gaze was absent, the preference to look at the target character seems to develop more slowly than in the condition the speaker's gaze was available. For the shift time window, the coefficients, *SEs*, *dfs*, and *t*-values as well as the *p*-values are shown in Table 1. No significant effect of speaker or agent on participants' fixations emerged. That implies that the two visual cues did not have an influence on participants' gaze behavior toward the NP2 referent (target) or the unmentioned third character competitor. What we do find though is a significant effect of time, i.e., for the 100*ms* time bins (*tiwi*) that we included as a factor  This indicates that participants' fixation behavior toward the target changed over the course of this time window.

Table 2 lists the coefficients, *SEs*, *dfs*, and *t*-values as well as the *p*-values for the second time window for Experiment 1. In this *NP2 time window*, participants looked more at the target character, which was mentioned by the speaker during this phase (see Table 2), than at the unmentioned competitor character. This is evident in the significant intercept ($p < 0.001$). Moreover, we find main effects of speaker as well as time bin (tiwi). Participants looked more at the target character when speaker gaze was available (vs. absent). Looks toward the target character increased as it was mentioned.

To sum up, virtual agent listener gaze seems to have no effect on participants' gaze behavior toward either target or competitor. The only visual cue that people used was speaker gaze.

### 2.3.2. Experiment 2

The only difference between the two experiments was that while the NP2 was audible in the first experiment, in Experiment 2, it was covered by pink noise. Figure 4B plots participants' fixation behavior in terms of gaze probability *log* ratios for the time between the onset of the spoken sentence until its offset. It illustrates that when only human speaker gaze was visible, participants already looked at the NP2 referent during the *shift time window* (which started at around 2,251 *ms*); their looks to the NP2 target declined again during that same region, only then to slightly rise again during the *NP2 time window* (beginning at around 3,181 *ms*). When both human speaker gaze and virtual agent listener gaze were present, participants tended to look at the competitor first but then started fixating the target character during the *shift time window*. These looks continued to increase during the NP2 region (this time window started at 4,612 *ms*). When either only the virtual agent or none of the two interlocutors was visible, people tended to look at the target character early in the *shift time window* (from shift onset until the onset of the NP2) and then directed their gaze to the competitor

3 The Eyelink 1,000 eye tracker has a blink/occlusion recovery of $M < 1.8$ *ms*, $SD < 0.6$ *ms* at a sampling rate of 1,000 Hz (SR Research, 2010).

4 See also Arai et al. (2007, here esp. Section 2.2.2) and Baayen (2008).

5 In all analyzes, convergence was achieved for Model 1 under R version 3.3.3 (R Core Team, 2017) using version 1.1–12 of the package lme4 (Bates et al., 2015) by the first author, the results for which will be reported in the following sections. Note that in some other setups, convergence was achieved only after simplifying the model by removing the interaction from the by-items random component. However, the results were always comparable to the ones reported here.

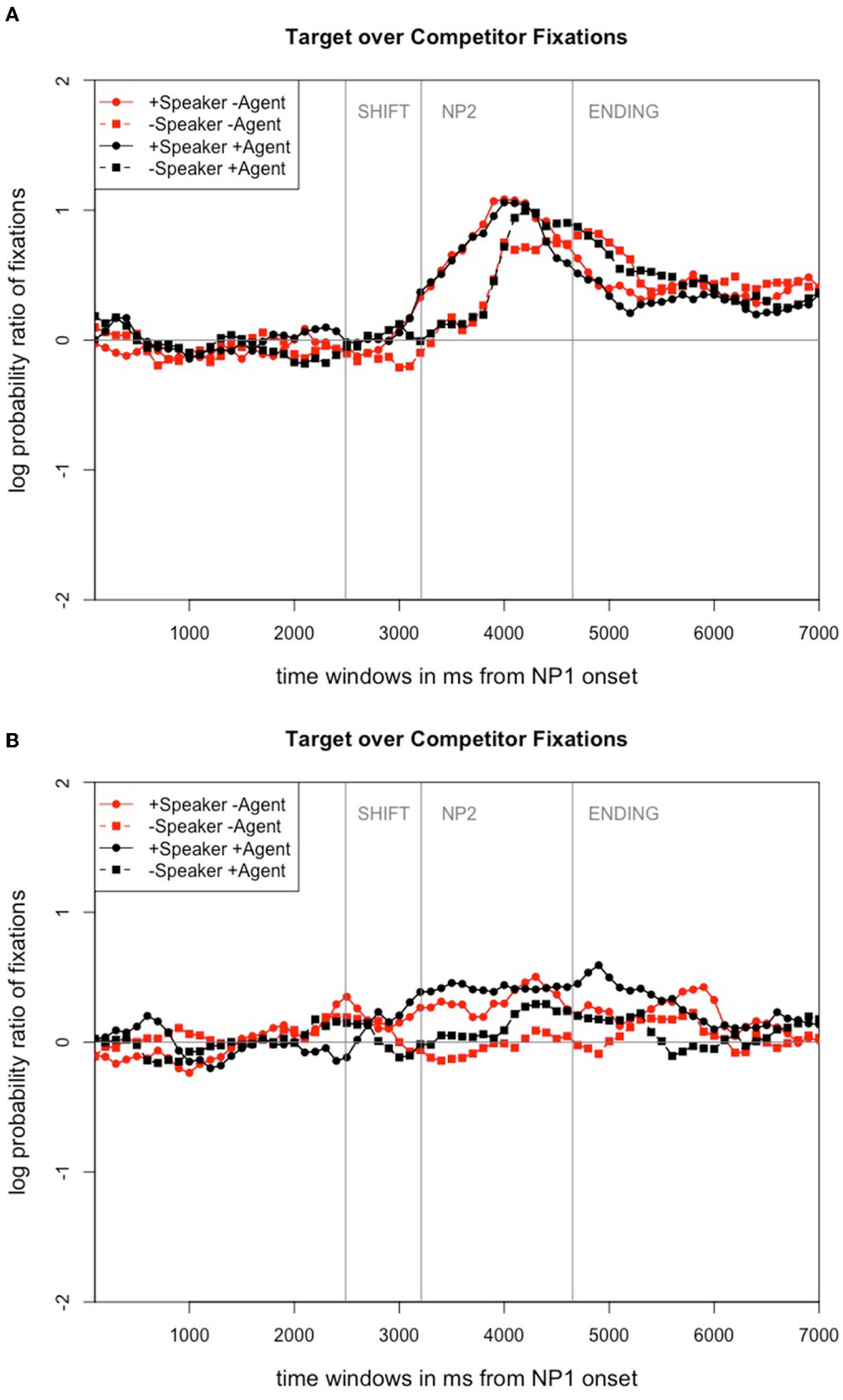


FIGURE 4
Time course graphs (in *ms*) from NP1 onset until ending offset. **(A)** Time course graph for Experiment 1. Adapted from Nunnemann (2021, p. 89). **(B)** Time course graph for Experiment 2. Adapted from Nunnemann (2021, p. 112).

character at the end of this time window. When neither agent nor human were visible, participants inspected the NP2 target. When neither agent nor human were visible, participants inspected the NP2 target.

The coefficients, *SE*s, *df*s, and *t*-values, as well as *p*-values, for the model of the shift time window of the experiment are shown in Table 3. Only the factor speaker showed a trend toward significance ($p = 0.051$). This implies a slight tendency for participants to

TABLE 1 Coefficients, *SE*s, *df*s, *t*-values, and *p*-values for the optimal model of log ratios of target fixations for the shift time window in Experiment 1.

| Fixed effects: | | | | | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. error | *df* | *t*-value | $Pr(>\|t\|)$ | |
| (Intercept) | $6.968e-02$ | $1.133e-01$ | $4.070e+01$ | 0.615 | 0.5421 | |
| cspeaker | $-3.319e-03$ | $1.978e-01$ | $2.900e+01$ | -0.017 | 0.9867 | |
| cagent | $1.037e-01$ | $2.099e-01$ | $3.930e+01$ | 0.494 | 0.6239 | |
| ctiwi | $-4.696e-02$ | $1.926e-02$ | $2.736e+03$ | -2.439 | 0.0148 | * |
| cspeaker:cagent | $-1.944e-02$ | $4.118e-01$ | $3.770e+01$ | -0.047 | 0.9626 | |
| cspeaker:ctiwi | $5.489e-02$ | $3.849e-02$ | $2.738e+03$ | 1.426 | 0.1540 | |
| cagent:ctiwi | $2.834e-03$ | $3.855e-02$ | $2.731e+03$ | 0.074 | 0.9414 | |
| cspeaker:cagent:ctiwi | $7.680e-02$ | $7.705e-02$ | $2.735e+03$ | -0.997 | 0.3190 | |

*$p < 0.05$.

TABLE 2 Coefficients, *SE*s, *df*s, *t*-values, and *p*-values for the optimal model of logratios of target fixations for the NP2 time window in Experiment 1.

| Fixed effects: | | | | | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. error | *df* | *t*-value | $Pr(>\|t\|)$ | |
| (Intercept) | $4.992e-01$ | $1.119e-01$ | $4.200e+01$ | 4.459 | $6.10e-0$ | *** |
| cspeaker | $8.758e-01$ | $2.929e-01$ | $4.200e+01$ | 2.990 | 0.00466 | ** |
| cagent | $4.511e-02$ | $2.082e-01$ | $4.100e+01$ | 0.217 | 0.82955 | |
| ctiwi | $7.885e-02$ | $1.587e-02$ | $3.226e+03$ | 4.968 | $7.13e-07$ | *** |
| cspeaker:cagent | $-1.506e-01$ | $3.963e-01$ | $4.200e+01$ | -0.380 | 0.70580 | |
| cspeaker:ctiwi | $5.909e-02$ | $3.175e-02$ | $3.228e+03$ | 1.861 | 0.06279 | ◦ |
| cagent:ctiwi | $2.418e-03$ | $3.175e-02$ | $3.227e+03$ | 0.076 | 0.93930 | |
| cspeaker:cagent:ctiwi | $-4.967e-02$ | $6.352e-02$ | $3.224e+03$ | -0.782 | 0.43433 | |

◦$p < 0.1$, **$p < 0.01$, ***$p < 0.001$.

look more at the target of the (unintelligible) NP2 when they could exploit human speaker gaze than in any other condition. These results are in contrast to our findings from Experiment 1, for which a significant effect of time emerged.

For the analysis of the NP2 time window, Table 4 lists the coefficients, the *SE*s, *df*s, and *t*-values as well as the *p*-values. Here, we find a significant effect of speaker. This indicates that participants looked more at the target character whenever speaker gaze was available. In contrast, agent listener gaze did not have any effect on participants' gaze behavior, although in this time window, the virtual agent listener also had shifted its gaze toward the target character. These findings replicate the results from Experiment 1. Moreover, the significant intercept indicates that participants looked more toward the target character than the competitor, although the NP2 referent was not audible throughout Experiment 2. Here again, we replicated our findings that participants looked more to the target during the later *NP2 time window* when the speaker's gaze cues were present.

### 2.3.3. Experiment 3

In Experiment 3, the NP2 was audible again and the agent listener inspected the target referent about 400 *ms* before the human speaker. The time course graph for Experiment 3 depicts the log probability ratio of fixations toward the target over the competitor for all four conditions (see Figure 5A). Here, we can see that participants seemed to look at the competitor more often than at the target character during the agent shift. The graphs for the agent-only condition and the no human speaker/no agent listener condition only differ slightly from each other. Participants looked more at the competitor during both *shift time windows* (that of the agent listener and that of the human speaker from the onset of their gaze shifts). This gaze behavior only changed during the *NP2 time window* when the human speaker named the target character. In those two conditions in which speaker gaze was available as a cue, participants already started to fixate the target during the *speaker shift time window*. Thus, the graphical inspection of the eye-tracking data in the time course graph already suggests that speaker

TABLE 3 Coefficients, *SE*s, *df*s, *t*-values, and *p*-values for the optimal model of logratios of target fixations for the shift time window for Experiment 2.

| Fixed effects: | | | | | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. error | *df* | *t*-value | $Pr(>\|t\|)$ | |
| (Intercept) | 0.17621 | 0.10602 | 36.90000 | 1.662 | 0.1050 | |
| cspeaker | 0.43013 | 0.21348 | 36.90000 | 2.015 | 0.0513 | ° |
| cagent | -0.04748 | 0.24075 | 38.40000 | -0.197 | 0.8447 | |
| ctiwi | -0.02061 | 0.01992 | 2419.70000 | -1.035 | 0.3009 | |
| cspeaker:cagent | 0.06818 | 0.45795 | 38.30000 | 0.149 | 0.8824 | |
| cspeaker:ctiwi | 0.02563 | 0.03987 | 2419.20000 | 0.643 | 0.5204 | |
| cagent:ctiwi | 0.01043 | 0.03982 | 2419.90000 | 0.262 | 0.7933 | |
| cspeaker:cagent:ctiwi | 0.03435 | 0.07969 | 2419.40000 | 0.431 | 0.6664 | |

$^{\circ}p < 0.1$.

TABLE 4 Coefficients, *SE*s, *df*s, *t*-values, and *p*-values for the optimal model of logratios of target fixations for the NP2 time window for Experiment 2.

| Fixed effects: | | | | | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. error | *df* | *t*-value | $Pr(>\|t\|)$ | |
| (Intercept) | $3.434e-01$ | $1.640e-01$ | $4.010e+01$ | 2.094 | 0.04263 | * |
| cspeaker | $7.044e-01$ | $2.370e-01$ | $4.010e+01$ | 2.973 | 0.00498 | ** |
| cagent | $2.663e-01$ | $3.025e-01$ | $4.220e+01$ | 0.880 | 0.38364 | |
| ctiwi | $1.484e-02$ | $1.760e-02$ | $2.569e+03$ | 0.843 | 0.39904 | |
| cspeaker:cagent | $3.506e-01$ | $5.168e-01$ | $3.720e+01$ | 0.678 | 0.50168 | |
| cspeaker:ctiwi | $9.146e-03$ | $3.524e-02$ | $2.572e+03$ | 0.260 | 0.79526 | |
| cagent:ctiwi | $3.709e-02$ | $3.521e-02$ | $2.568e+03$ | 1.053 | 0.29225 | |
| cspeaker:cagent:ctiwi | $1.298e-01$ | $7.049e-02$ | $2.570e+03$ | -1.841 | 0.06573 | |

$^{*}p < 0.05$, $^{**}p < 0.01$.

gaze cues are more helpful as they result in earlier and more looks to the NP2 target character.

Table 5 lists coefficients, *SE*s, *df*s, and *t*-values, as well as *p*-values, for the analysis of the shift time window. Speaker gaze had a significant effect on participant eye movement behavior: They looked more toward the target when speaker gaze was visible than when it was absent. Furthermore, we found a significant effect of time window. This indicates that participants looked more to the target character over time. Also the interaction between speaker and time bin is significant. This implies that fixations toward the correct NP2 referent increased over time whenever the human speaker was present. In the shift time window for Experiment 3, this holds also true for those conditions in which the virtual agent's gaze was available. The interaction between time bin and agent is significant as well. This could be explained by the fact that the cues from virtual agent listener gaze preceded those from human speaker gaze in Experiment 3.

The second time window of interest in Experiment 3 is the *NP2 time window* for which the results are detailed in Table 6. Overall, for this model, there is a significant intercept for which the results mean that participants fixated the NP2 target character more than the competitor. Furthermore, the factor speaker is significant suggesting that the presence of the human speaker's gaze had a systematic influence on participants' gaze behavior toward the NP2 referent. Moreover, the factor tiwi was significant. Thus, participants gaze more toward the target character of the NP2 as more time passes.

### 2.3.4. Experiment 4

Figure 5B displays participants' looks toward the target character over those toward the competitor in Experiment 4 in which the agent listener inspected the NP2 target referent about 400 *ms* before the human speaker, and the NP2 was inaudible. The visual inspection reveals that people only started looking more toward the NP2 target character toward the end of the *NP2 time window* which started when the pink noise sets in. But this only holds true for the three conditions in which at least one type of gaze cue was visible. When neither the human speaker nor the virtual agent listener were

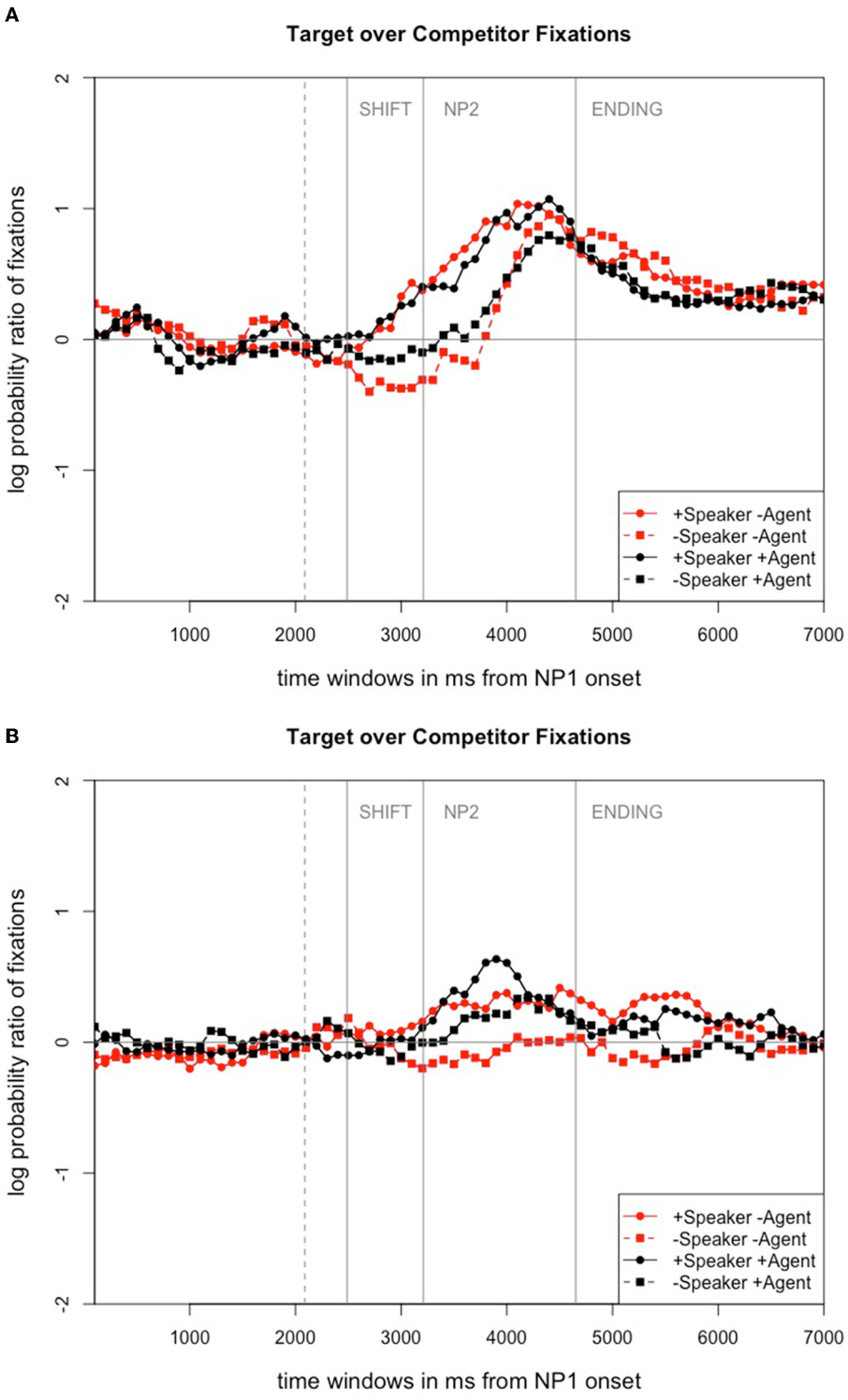


FIGURE 5
Time course graphs (in *ms*) from NP1 onset until ending offset for the experiments with altered agent listener gaze. The dotted gray line marks the earlier shift onset for the virtual agent. **(A)** Time course graph for Experiment 3. Adapted from Nunnemann (2021, p. 135). **(B)** Time course graph for Experiment 4. Adapted from Nunnemann (2021, p. 156).

present, people even looked slightly more to the unmentioned competitor than to the NP2 target referent during the *NP2 time window*.

Table 7 lists an overview of coefficients, *SE*s, *df*s, and *t*-values, as well as *p*-values, for the fixed effects of the shift time window. No significant main effects were found.

TABLE 5 Coefficients, *SE*s, *df*s, *t*-values, and *p*-values for the optimal model of logratios of target fixations for the shift time window in Experiment 3.

| Fixed effects: | | | | | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. error | *df* | *t*-value | $Pr(>\|t\|)$ | |
| (Intercept) | $1.896e-03$ | $1.173e-01$ | $4.140e+01$ | 0.016 | 0.98718 | |
| cspeaker | $4.683e-01$ | $1.896e-01$ | $3.220e+01$ | 2.470 | 0.01901 | * |
| cagent | $2.728e-01$ | $2.384e-01$ | $3.260e+01$ | 1.144 | 0.26078 | |
| ctiwi | $-4.597e-02$ | $1.824e-02$ | $2.877e+03$ | -2.520 | 0.01180 | * |
| cspeaker:cagent | $-1.172e-01$ | $3.771e-01$ | $3.860e+01$ | -0.311 | 0.75755 | |
| cspeaker:ctiwi | $7.499e-02$ | $3.652e-02$ | $2.876e+03$ | 2.053 | 0.04012 | * |
| cagent:ctiwi | $1.045e-01$ | $3.653e-02$ | $2.875e+03$ | 2.860 | 0.00426 | ** |
| cspeaker:cagent:ctiwi | $-9.566e-02$ | $7.306e-02$ | $2.876e+03$ | -1.309 | 0.19053 | |

$^{*}p < 0.05$, $^{**}p < 0.01$.

TABLE 6 Coefficients, *SE*s, *df*s, *t*-values, and *p*-values for the optimal model of logratios of target fixations for the NP2 time window in Experiment 3.

| Fixed effects: | | | | | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. error | *df* | *t*-value | $Pr(>\|t\|)$ | |
| (Intercept) | 0.28697 | 0.12166 | 39.00000 | 2.359 | 0.0234 | * |
| cspeaker | 0.99203 | 0.22416 | 40.00000 | 4.425 | $7.25e-05$ | *** |
| cagent | 0.17260 | 0.27400 | 43.00000 | 0.630 | 0.5321 | |
| ctiwi | 0.07397 | 0.01582 | 3297.00000 | 4.676 | $3.04e-06$ | *** |
| cspeaker:cagent | -0.67092 | 0.40021 | 39.00000 | -1.676 | 0.1017 | |
| cspeaker:ctiwi | 0.01838 | 0.03162 | 3302.00000 | 0.581 | 0.5610 | |
| cagent:ctiwi | -0.04312 | 0.03167 | 3297.00000 | -1.362 | 0.1733 | |
| cspeaker:cagent:ctiwi | 0.06868 | 0.06332 | 3298.00000 | 1.085 | 0.2782 | |

$^{*}p < 0.05$, $^{***}p < 0.001$.

TABLE 7 Coefficients, *SE*s, *df*s, *t*-values, and *p*-values for the optimal model of logratios of target fixations for the shift time window in Experiment 4, by participants.

| Fixed effects: | | | | | |
|---|---|---|---|---|---|
| | Estimate | Std. error | *df* | *t*-value | $Pr(>\|t\|)$ |
| (Intercept) | $1.900e-02$ | $1.147e-01$ | $3.340e+01$ | 0.166 | 0.869 |
| cspeaker | $1.967e-01$ | $2.498e-01$ | $3.560e+01$ | 0.788 | 0.436 |
| cagent | $3.073e-02$ | $2.639e-01$ | $3.810e+01$ | 0.116 | 0.908 |
| ctiwi | $1.488e-02$ | $2.067e-02$ | $2.188e+03$ | 0.720 | 0.472 |
| cspeaker:cagent | $-1.586e-02$ | $4.452e-01$ | $3.580e+01$ | -0.036 | 0.972 |
| cspeaker:ctiwi | $1.248e-02$ | $4.160e-02$ | $2.188e+03$ | 0.300 | 0.764 |
| cagent:ctiwi | $-1.966e-03$ | $4.136e-02$ | $2.189e+03$ | -0.048 | 0.962 |
| cspeaker:cagent:ctiwi | $-4.241e-02$ | $8.313e-02$ | $2.194e+03$ | -0.510 | 0.610 |

TABLE 8 Coefficients, *SE*s, *df*s, *t*-values, and *p*-values for the optimal model of logratios of target fixations for the NP2 time window in Experiment 4.

| Fixed effects: | | | | | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. error | *df* | *t*-value | $Pr(>\|t\|)$ | |
| (Intercept) | 0.39763 | 0.15144 | 39.00000 | 2.626 | 0.01229 | * |
| cspeaker | 0.68114 | 0.24615 | 35.90000 | 2.767 | 0.00888 | ** |
| cagent | 0.38724 | 0.28930 | 42.20000 | 1.339 | 0.18788 | |
| ctiwi | 0.03642 | 0.01888 | 2216.40000 | 1.929 | 0.05386 | ◦ |
| cspeaker:cagent | -0.24686 | 0.50867 | 39.10000 | -0.485 | 0.63017 | |
| cspeaker:ctiwi | 0.02426 | 0.03818 | 2221.20000 | 0.635 | 0.52524 | |
| cagent:ctiwi | 0.05992 | 0.03777 | 2215.60000 | 1.586 | 0.11278 | |
| cspeaker:cagent:ctiwi | 0.08592 | 0.07637 | 2221.00000 | 1.125 | 0.26066 | |

$^{\circ}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$.

The *NP2 time window* was overlaid with pink noise. But both interlocutors, i.e., the human speaker and the virtual agent listener would still fixate the target in those conditions in which they—and thus their gaze shifts—were visible. For the *NP2 time window*, the results are summarized in Table 8. First of all, there was a significant intercept. Overall, participants fixated the NP2 target character more than the competitor throughout the *NP2 time window*. In addition to that, an effect of speaker emerged: The presence of speaker gaze influenced participants' gaze behavior toward the NP2 referent. They looked significantly more to the NP2 referent when the speaker was visible than when she was absent. The human speaker's gaze likely facilitated detection of the target character, as the utterance identifying it was overlaid with pink noise.

## 3. General discussion

In four eye tracking experiments, we investigated the effects of human speaker and virtual agent listener gaze on spoken language comprehension. We wanted to know how the simultaneous presence of the two different types of gaze affected people's fixation behavior as they listened to utterances and inspected related videos. We were also interested in whether a virtual agent's gaze guides comprehenders' gaze in a similar fashion as a human speaker's gaze.

In four eye tracking experiments, we maximally contrasted human speaker and agent listener gazes and their effects on eye movements. For details on further outcomes, see Nunnemann (2021). In the present study, we reported the effects of two manipulations: the presence or absence of a human speaker's gaze and the presence or absence of a virtual agent listener's gaze. Common to all four experiments was that participants watched videos in which the human speaker and the virtual agent listener inspected a static scene with three clearly identifiable characters (e.g., a millionaire, a waiter, and a saxophone player). In Experiments 1 and 3, participants heard a sentence—uttered by the human speaker—that described an interaction between two of the three characters, e.g., *Der Kellner beglückwünscht den Millionär* ("The waiter congratulates the millionaire."). In Experiments 2 and 4, the NP2 was overlaid with pink noise, which made it inaudible. Thus, in those latter experiments, participants had to exclusively rely on human speaker/agent listener gaze cues to identify the correct NP2 referent. Another important difference between the experiments was the timing of the virtual agent's listener gaze. While in Experiments 1 and 2, the agent Billie displayed typical listener behavior by following human speaker gaze and fixating the target referent about 400 *ms* later than the speaker, in Experiments 3 and 4, the virtual agent gazed at the referent 400 *ms* before the speaker.

The main finding from all four experiments is that people do exploit a human speaker's gaze. But whereas our eye tracking results revealed that in Experiments 1, 2, and 4, participants only looked at the NP2 target character during the *NP2 time window*, which began when the human speaker started to utter the NP2 referent, participants in Experiment 3 already fixated the target character in the earlier *shift time window*, when the speaker shifted her head to fixate the target. Thus, these findings indicate that in the first two experiments, in which the speaker gazed at the NP2 referent 200 *ms* before she mentioned it, participants gazed at the target when the human speaker mentioned it. Only in Experiment 3, in which the virtual agent looked at the NP2 referent before the speaker shifted her gaze, participants already started looking at the target while the human speaker shifted her gaze. In Experiment 4, in which the NP2 is inaudible, they only start looking at the NP2 referent during the second time window—even though agent gaze identified the NP2 target before human gaze. Thus, in all four studies only human speaker gaze proved to be helpful.

The presented experiments have provided further evidence for the impact of referential gaze on situated language comprehension. First, we replicated earlier findings that people benefit from the gaze cues of a human speaker to identify a temporarily ambiguous target referent. In one of our own experiments—Experiment 3, where the agent gazed earlier at the NP2 target than the speaker—participants were even able to anticipate the correct NP2 referent already while the human speaker shifted her gaze toward it but before starting to name it. However, in contrast to the human speaker's gaze, which proved to be beneficial already while it was perceived, participants seemed to not react in their eye-movements to the virtual agent's

gaze—perhaps because of its virtual nature or perhaps because the agent always had the function of a listener and not a speaker. The gaze of the virtual agent listener seemed to be completely ignored except for an increase in looks to the target character over time if the agent's gaze was present (as corroborated by an interaction between agent gaze and time bin).

With these results, we replicated major findings from the field that showed that people do exploit a speaker's gaze in situated language comprehension. Thus, in line with the studies by Kreysa and Knoeferle (2011b), Knoeferle and Kreysa (2012), and Kreysa et al. (2018), for instance, we found that people are able to anticipate a referent before it is fully mentioned by a speaker and that they also do gaze more at the target when the speaker's gaze shift toward the referent was visible. Furthermore, a speaker's gaze helps to resolve temporary ambiguities in language understanding (Hanna and Brennan, 2007). A possible explanation for such a selective use of gaze cues—as also observed by Raidt et al. (2005)—comes from recent studies by Sekicki and Staudte (2018) as well as Jachmann et al. (2019). The latter authors argue that the exploitation of a speaker's gaze in situated language comprehension does not increase cognitive load unless the gaze cues and the utterances are incongruent. Although in our four experiments the agent gaze cues were always congruent with the linguistic input, participants largely ignored the agent's fixations. Perhaps watching two characters while integrating their gaze during full sentence comprehension is cognitively demanding, and this increased load might interfere with any beneficial effects of a gaze cue that is perhaps not perceived as sufficiently human-like or for which the interlocutor did not have the right communicative (speaker) role. This might be a possible explanation for this finding. In a study by Kulms and Kopp (2013), participants do not fully rely on agent gaze in multiple task scenarios. In our first two experiments, the virtual agent's gaze always followed the gaze of the speaker at the pace of a typical listener. The speaker already started uttering the NP2 when the agent fixated the target, and thus, at least in conditions in which both interlocutors were present, the agent's gaze might have been perceived as redundant.

We also had speculated that people might ignore agent gaze it was artificial or it might not be perceived in a similar way as a speaker's gaze. This line of argument is contradicted by earlier research that found that artificial gaze cues can generally be exploited (e.g., Staudte and Crocker, 2009; Martinez et al., 2010; Andrist et al., 2014; Mwangi et al., 2018). They even can be rather rudimentary, as the study by Boucher et al. (2012) showed. Whereas, these studies only provided evidence that the gaze cues from a robotic or virtual agent can be used in general, Rehm and André (2005) provided evidence that people in a multiparty conversation paid more attention to an agent interlocutor than to a fellow human interlocutor. They reasoned that this observation was due to the novelty effect of the agent. Yu et al. (2012) describe a similar effect in their results, as people who interacted with a virtual agent adapted more to it than when they had the same interaction with a human interlocutor. Given these findings in the literature, our findings that people did not pay any attention to the virtual agent and even ignored it despite its usefulness seem surprising. But then again, Rehm and André (2005) observed that the novelty effect of the agent disappeared when it was not the speaker but the addressee in the conversation. The communicative role of the virtual agent seems to have an influence on its perception. To disentangle whether agent gaze cues were not exploited due to the virtual agent being the listener of the interaction, it was an artificial agent, or even due to a combination of both aspects (see Rehm and André, 2005), the next logical step would be to swap the communicative roles. That means the human would be the passive listener, while the agent would take over the role of the speaker. In addition, future studies should replicate our findings when switching the respective positions of the human and the virtual agent on the screen. We refrained from adding this further complexity to the design of the current studies, but there is of course a possibility that an interlocutor on the right of the screen (in the current studies, always the human speaker) might receive more attention than one on the left of the screen (i.e., the virtual agent).

Even though there are still aspects that we have not touched upon, our series of studies and their outcomes have contributed to better understanding the comprehension of situated language and the impact that nonverbal cues, such as speaker and listener gaze, have on these processes. The results and insights from our studies can help to inform the creation of communicative virtual agents, employed for instance as teachers, tutors, or guides. Another example of an area where artificial agents might be applied in the near future is the field of therapy and assistance. Applications involving virtual agents and humans on a display may not only require these artificial interlocutors to understand and produce spoken language but also to display appropriate nonverbal behavior.

Here, especially referential gaze is a very important aspect in spoken language usage, as it can—as our studies have shown for speaker gaze—facilitate language comprehension. Although the gaze of a virtual agent listener did not have a facilitating effect in our experiments, it is still crucial to also gain insight into the mechanisms underlying the processing of virtual agent gaze. Thus, our experiments contributed to the ongoing basic research in this area. They can also contribute to the development of more elaborate models of situated language comprehension by adding information on gaze cues. These could then be implemented in virtual agents. In future, this could enable virtual agents to act as competent communication partners that can interact with humans.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation. Alternatively, the datasets and analysis scripts are available on OSF via the following link: https://osf.io/gp9cy/?view_only=a1cd0113218045caaf4ebbd636bdcf3c.

## Ethics statement

The studies involving human participants were reviewed and approved by Ethics Committee of

Bielefeld University (EUB). The patients/participants provided their written informed consent to participate in this study.

## Author contributions

EN, HK, and PK contributed to the design of the studies. EN conducted the experiments, analyzed the data, and provided a first draft of the manuscript. HK and PK commented on the draft and provided feedback. The authors thereby agree to be accountable for the content of the article. All authors contributed to the article and approved the submitted version.

## Funding

This research was funded by grants from the Cluster of Excellence Cognitive Interaction Technology CITEC (EXC 277), Bielefeld University. The article processing charge was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 491192747 and the Open Access Publication Fund of Humboldt-Universität zu Berlin.

## Acknowledgments

The authors would like to thank the working group Sociable Agents at CITEC for providing the virtual agent for the described studies. Moreover, we would like to thank Dr. Kirsten Bergmann for her advice on the setup of the agent. The content of the present article is based upon a previously published doctoral thesis (Nunnemann, 2021) at Bielefeld University, Germany.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## References

Andrist, S., Mutlu, B., and Gleicher, M. (2013). "Conversational gaze aversion for virtual agents," in *International Workshop on Intelligent Virtual Agents* (Springer), 249–262. doi: 10.1007/978-3-642-40415-3_22

Andrist, S., Pejsa, T., Mutlu, B., and Gleicher, M. (2012). "A head-eye coordination model for animating gaze shifts of virtual characters," in *Gaze-In '12: Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction* (ACM), 1–6. doi: 10.1145/2401836.2401840

Andrist, S., Tan, X. Z., Gleicher, M., and Mutlu, B. (2014). "Conversational gaze aversion for humanlike robots," in *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE), 25–32. doi: 10.1145/2559636.2559666

Arai, M., Van Gompel, R. P. G., and Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cogn. Psychol.* 54, 218–250. doi: 10.1016/j.cogpsych.2006.07.001

Argyle, M., and Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge: Cambridge University Press.

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bee, N., and André, E. (2008). "Cultural gaze behavior to improve the appearance of virtual agents," in *IUI Workshop on Enculturating Interfaces (ECI)*, 1–5.

Boucher, J.-D., Pattacini, U., Lelong, A., Bailly, G., Elisei, F., Fagel, S., et al. (2012). I reach faster when i see you look: gaze effects in human-human and human-robot face-to-face cooperation. *Front. Neurorobot.* 6, 3. doi: 10.3389/fnbot.2012.00003

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., and Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*. 106, 1465–1477. doi: 10.1016/j.cognition.2007.05.012

Carpenter, P. A., and Just, M. A. (1975). Sentence comprehension: a psycholinguistic processing model of verification. *Psychol. Rev.* 82, 45. doi: 10.1037/h0076248

Courgeon, M., Rautureau, G., Martin, J.-C., and Grynszpan, O. (2014). Joint attention simulation using eye-tracking and virtual humans. *IEEE Trans. Affect. Comput.* 5, 238–250. doi: 10.1109/TAFFC.2014.2335740

Fischer, B., and Breitmeyer, B. (1987). Mechanisms of visual attention revealed by saccadic eye movements. *Neuropsychologia* 25, 73–83. doi: 10.1016/0028-3932(87)90044-3

Hanna, J. E., and Brennan, S. E. (2007). Speakers eye gaze disambiguates referring expressions early during face-to-face conversation. *J. Mem. Lang.* 57, 596–615. doi: 10.1016/j.jml.2007.01.008

Heylen, D., Nijholt, A., and Poel, M. (2007). "Generating nonverbal signals for a sensitive artificial listener," in *Verbal and Nonverbal Communication Behaviours. Lecture Notes in Computer Science, volume 4775*, eds A. Esposito, M. Faundez-Zanuy, E. Keller, and M. Marinaro (Springer), 264–274. doi: 10.1007/978-3-540-76442-7_23

Jachmann, T. K., Drenhaus, H., Staudte, M., and Crocker, M. W. (2019). Influence of speakers gaze on situated language comprehension: evidence from event-related potentials. *Brain Cogn.* 135, 1–12. doi: 10.1016/j.bandc.2019.05.009

Johnson, W. L., Rickel, J. W., and Lester, J. C. (2000). Animated pedagogical agents: face-to-face interaction in interactive learning environments. *Int. J. Artif. Intell. Educ.* 11, 47–78.

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

Kreysa, H., and Knoeferle, P. (2011a). "Effects of speaker gaze on spoken language comprehension: task matters," in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society, volume 33*, eds L. Carlson, C. Hoelscher, and T. Shipley (Cognitive Science Society). Available online at: https://escholarship.org/uc/item/80q806kh

Kreysa, H., and Knoeferle, P. (2011b). "Peripheral speaker gaze facilitates spoken language comprehension: Syntactic structuring and thematic role assignment in German," in *Proceedings of the European Conference on Cognitive Science*, eds B. Kokinov, A. Karmiloff-Smith, and N. Nersessian (New Bulgarian University Press).

Kreysa, H., Nunnemann, E. M., and Knoeferle, P. (2018). Distinct effects of different visual cues on sentence comprehension and later recall: the case of speaker gaze versus depicted actions. *Acta Psychol.* 188, 220–229. doi: 10.1016/j.actpsy.2018.05.001

Kulms, P., and Kopp, S. (2013). "Using virtual agents to guide attention in multi-task scenarios," in *International Workshop on Intelligent Virtual Agents. IVA 2013. Lecture Notes in Computer Science, volume 8108* (Springer), 295–302. doi: 10.1007/978-3-642-40415-3_26

Lance, B. J., and Marsella, S. (2010). The expressive gaze model: Using gaze to express emotion. *IEEE Comput. Graphics Appl.* 30, 62–73. doi: 10.1109/MCG.2010.43

Lee, J., and Marsella, S. (2006). "Nonverbal behavior generator for embodied conversational agents.," in *International Workshop on Intelligent Virtual Agents.*

*IVA 2006. Lecture Notes in Computer Science, volume 4133* (Springer), 243–255. doi: 10.1007/11821830_20

Maatman, R., Gratch, J., and Marsella, S. (2005). "Natural behavior of a listening agent," in *Intelligent Virtual Agents. IVA 2005. Lecture Notes in Computer Science, volume 3661*, eds T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, and T. Rist (Berlin: Springer), 25–36.

Martinez, S., Sloan, R. S., Szymkowiak, A., and Scott-Brown, K. (2010). "Using virtual agents to cue observer attention," in *CONTENT 2010: The Second International Conference on Creative Content Technologies*, 7–12.

Mwangi, E., Barakova, E. I., Díaz-Boladeras, M., Mallofré, A. C., and Rauterberg, M. (2018). Directing attention through gaze hints improves task solving in human-humanoid interaction. *Int. J. Soc. Rob.* 10, 343–355. doi: 10.1007/s12369-018-0473-8

Nunnemann, E. M. (2021). *The influence of referential gaze on spoken language comprehension: Human speaker vs*. virtual agent *listener gaze* (Ph.D. thesis). Universität Bielefeld. doi: 10.4119/unibi/2964479

Pfeiffer-Leßmann, N., and Wachsmuth, I. (2008). "Toward alignment with a virtual human-Achieving joint attention," in *KI 2008: Advances in Artificial Intelligence. KI 2008. Lecture Notes in Computer Science*, eds A. R. Dengel, K. Berns, T. M. Breuel, F. Bomarius, and T. R. Roth-Berghofer (Springer). doi: 10.1007/978-3-540-85845-4_36

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Raidt, S., Elisei, F., and Bailly, G. (2005). "Face-to-face interaction with a conversational agent: Eye-gaze and deixis," in *International Conference on Autonomous Agents and Multiagent Systems*, 17–22. Available online at: https://hal.science/hal-00419299/en

Rehm, M., and André, E. (2005). "Where do they look? Gaze behaviour of multiple users interacting with an embodied conversational agent," in *Intelligent Virtual Agents. IVA 2005, volume 3661*, eds T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, and P. Olivier (Springer), 241–252. doi: 10.1007/11550617_21

Richardson, J. T. (2018). The use of latin-square designs in educational and psychological research. *Educ. Res. Rev.* 24, 84–97. doi: 10.1016/j.edurev.2018.03.003

Sekicki, M., and Staudte, M. (2018). Eyell help you out! How the gaze cue reduces the cognitive load required for reference processing. *Cogn. Sci*. 42, 2418–2458. doi: 10.1111/cogs,.12682

SR Research (2010). *Eyelink 1000 user's manual*. version 1.5.2.

Staudte, M., and Crocker, M. W. (2009). "Visual attention in spoken human-robot interaction," in *Proceedings of the 4th ACM/IEEE international conference on Human-Robot Interaction (HRI)* (La Jolla, CA: IEEE), 77–84.

Staudte, M., and Crocker, M. W. (2011). Investigating joint attention mechanisms through spoken human-robot interaction. *Cognition* 120, 268–291. doi: 10.1016/j.cognition.2011.05.005

Steptoe, W., Oyekoya, O., Murgia, A., Wolff, R., Rae, J., Guimaraes, E., et al. (2009). "Eye tracking for avatar eye gaze control during object-focused multiparty interaction in immersive collaborative virtual environments," in *2009 IEEE Virtual Reality Conference* (Lafayette, LA: IEEE), 83–90.

Wang, N., and Gratch, J. (2010). "Don't just stare at me!" in *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1241–1250. doi: 10.1145/1753326.1753513

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). "Elan: a professional framework for multimodality research," in *5th International Conference on Language Resources and Evaluation (LREC 2006)*, *Vol. 2006*, 1556–1559. Available online at: https://hdl.handle.net/11858/00-001M-0000-0013-1E7E-4

Yu, C., Schermerhorn, P., and Scheutz, M. (2012). Adaptive eye gaze patterns in interactions with human and artificial agents. *ACM Trans. Interact. Intell. Syst.* 1, 1–25. doi: 10.1145/2070719.2070726