Cognition 133 (2014) 317–328

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

The influence of speaker gaze on listener comprehension: Contrasting visual versus intentional accounts

Maria Staudte [a,*], Matthew W. Crocker [a], Alexis Heloir [b,c], Michael Kipp [d]

[a] Department of Computational Linguistics, Saarland University, Germany
[b] DFKI – MMCI, Saarland University, Germany
[c] LAMIH-UMR CNRS 8201, Le Mont Houy, Univ. Valenciennes, France
[d] Computer Science, Augsburg University of Applied Sciences, Germany

ARTICLE INFO

Article history:
Received 8 February 2012
Revised 6 June 2014
Accepted 10 June 2014
Available online 1 August 2014

Keywords:
Joint attention
Gaze
Arrows
Visual attention shifts
Referential intention
Language comprehension

ABSTRACT

Previous research has shown that listeners follow speaker gaze to mentioned objects in a shared environment to ground referring expressions, both for human and robot speakers. What is less clear is whether the benefit of speaker gaze is due to the inference of referential intentions (Staudte and Crocker, 2011) or simply the (reflexive) shifts in visual attention. That is, is gaze special in how it affects simultaneous utterance comprehension? In four eye-tracking studies we directly contrast speech-aligned speaker gaze of a virtual agent with a non-gaze visual cue (arrow). Our findings show that both cues similarly direct listeners' attention and that listeners can benefit in utterance comprehension from both cues. Only when they are similarly precise, however, does this equality extend to incongruent cueing sequences: that is, even when the cue sequence does not match the concurrent sequence of spoken referents can listeners benefit from gaze as well as arrows. The results suggest that listeners are able to learn a counter-predictive mapping of both cues to the sequence of referents. Thus, gaze and arrows can in principle be applied with equal flexibility and efficiency during language comprehension.

1. Introduction

In face-to-face communication, the speaker's gaze to objects in a shared scene provides the listener with a visual cue to the speaker's focus of (visual) attention (Emery, 2000; Flom, Lee, & Muir, 2007). This potentially offers the listener valuable information to ground and disambiguate referring expressions, to hypothesize about the speaker's communicative intentions and goals and, thus, to facilitate comprehension (Hanna & Brennan, 2007). It is an open question, however, whether this functionality of speaker gaze results simply from its established ability to drive listeners' visual attention, as do other visual cues, or whether gaze uniquely expresses (referential) intentions.

More precisely, there are two levels on which a visual attention shift in response to a speaker's gaze may affect utterance processing: on a perceptual level, gaze-following may be considered as (reflexive) visuo-spatial orienting which increases the visual saliency of the particular target object and/or location in focus (Driver et al., 1999; Friesen & Kingstone, 1998; Langton & Bruce, 1999). On a cognitive level, gaze may *additionally* be understood as a cue to the speaker's referential intentions which elicits expectations about which referent would be mentioned next (Hanna & Brennan, 2007). These two levels have been dubbed the *Visual* and the *Intentional Account*, respectively (Staudte & Crocker, 2011). Whether the Intentional Account – and not the Visual Account alone – is necessary to explain such

* Corresponding author.
E-mail address: masta@coli.uni-saarland.de (M. Staudte).

http://dx.doi.org/10.1016/j.cognition.2014.06.003
0010-0277/

gaze effects on utterance comprehension, is still under debate. However, recent evidence provides converging support for the a view that gaze uniquely conveys intentions and mental states, above and beyond the pure attention shift that it also induces (Becchio, Bertone, & Castiello, 2008; Meltzoff, Brooks, Shon, & Rao, 2010; Staudte & Crocker, 2011).

Meltzoff et al. (2010), for instance, showed that infants who saw a robot previously engage in a social interaction with adults were more likely to follow the robot's gaze than those infants who had not had this experience. This result suggests that it is important for infants to recognize the robot as a social being, who perceives with its "eyes", in order to follow its gaze. Further, Staudte and Crocker (2011) synchronized gaze movements of a robot with its speech in a human-like manner. When played back in a video, these gaze movements were shown to be similarly useful to listeners for grounding and resolving spoken references as human gaze (Hanna & Brennan, 2007), even when preceding the respective verbal reference by several seconds (Staudte & Crocker, 2010). These findings suggest that gaze is interpreted (a) to be a socially relevant cue, and (b) with respect to a referential intention of the speaker which the listener maintains over time until realized (or until overridden by some other information as is probably the case in scenarios like the Human Simulation Project, Trueswell et al., unpublished). The critical question is whether all these results could have also been achieved if it had not been *gaze* directing participants' attention in each of these settings and circumstances but some other (potentially even coincidental) visual cue.

To date, few other on-line studies have investigated (speaker) gaze as a truly dynamic and embodied cue – rather than a static line drawing, for instance – and how this affects concurrent language processing of the listener. Critically, almost all of these studies (the Meltzoff study is one exception but does not include language comprehension) have only considered the congruence or credibility of gaze cues (e.g., Hanna & Brennan, 2007; Nappa, Wessel, McEldoon, Gleitman, & Trueswell, 2009; Staudte & Crocker, 2010, 2011). That is, to our knowledge it has not been investigated so far whether or not the observed effects of gaze on utterance comprehension are due to the elicited shifts in listeners' visual attention *per se* and can in principle be evoked by any other direction-giving cue, or whether they are indeed unique to gaze. Partly, this lack of evidence is due to the difficulty of evoking incongruent human gaze and speech, and partly it is due to the difficulty of comparing a human gesture or gaze cue with other, more artificial cues. One way to overcome these constraints is to employ an artificial agent who is, on the one hand, fully controllable in its behavior and, on the other hand, is likely more accepted when producing incongruent or atypical behavior. Further, when situated in virtual environments, fair comparison with other visual cues is facilitated.

Thus, to explore the hypothesis that speaker gaze is *uniquely* interpreted with respect to referential intentions, we directly compare the influence of agent gaze to a purely visual baseline cue by replacing the gaze movement of a virtual agent with an arrow. Specifically, we report evidence from four studies that, firstly, replicate and extend previous findings concerning the relevance of gaze cue order for comprehension (Experiment 1). Secondly, a baseline study showing arrow cues revealed that, while being more visually precise, such arrow cues were used more effectively and flexibly to support utterance comprehension (Experiments 2a and 2b). And finally, Experiment 3 shows that a visually precise gaze cue in a simplified scene can also be exploited in a flexible and efficient manner by the listener. Together, these findings suggest that gaze and arrows direct attention and visually highlight the cued objects in a similar way. However, speaker gaze may frequently be spatially imprecise, as was the case in Experiment 1, such that it is harder to exploit speaker gaze for utterance comprehension when the concurrent utterance does not match the cueing order. This disadvantage can be overcome when speaker gaze is as visually precise as the arrow baseline. Both cues can then be used similarly by listeners to infer and anticipate an upcoming verbal reference. Thus, the predictive effect of speaker gaze for a listener seems to be solely a (learnable) effect of cueing a given object at a given time which is *independent* of the potential intention or mental state attributed to the gazing speaker.

1.1. Reflexive and voluntary orienting to cues

Previous studies have shown that people reflexively follow stylized gaze cues and other direction-giving cues like arrows to a target location (e.g., Ristic, Friesen, & Kingstone, 2002). Whether gaze and arrow cues, for instance, elicit the same type of attention shift or whether gaze is in some way special is still under examination (Bayliss & Tipper, 2005; Ristic, Wright, & Kingstone, 2007; Tipples, 2008; Vaidya et al., 2011). Beyond the reflexive attention shifts mentioned above, people have further been shown to voluntarily orient towards symbolic cues when there is reason to consider these as useful (e.g., Posner, 1980). That is, when arrow cues are learned to be counter-predictive (cueing one direction but reliably predicting the target in another direction), they also trigger slightly delayed, voluntary attention shifts (Friesen, Ristic, & Kingstone, 2004; Tipples, 2008, or Hanna & Brennan, 2007, for gaze cues).

Thus, evidence suggests that reflexive, and potentially also voluntary, orienting applies to both gaze and arrows. Simultaneously, a large body of research has shown that gaze not only drives visual attention but that it further reveals complex mental states and even intentions (Baron-Cohen, Campbell, Karmiloff-Smith, Grant, & Walker, 1995; Meltzoff et al., 2010). It seems indeed plausible that a life-time of experiences has taught people that gaze can reveal somebody's beliefs, intentions, or emotions, and how useful this may be for communication (Tomasello & Carpenter, 2007). Thus, the motivation to follow gaze and effects thereof may well be special and unique when it comes to integration with concurrent language. In order to tease apart any effects of the (reflexive and voluntary) visual cueing function of gaze (cf. the Visual Account) and the elicited inference of referential intentions (Intentional Account), a baseline providing *only* the visual

cueing function is required. Arrow cues constitute such a baseline and potentially diverging effects on comprehension may shed light on whether and how gaze is processed beyond its well-studied visual cueing function. Such a comparison may further reveal to what extent the cue (gaze versus arrow) that elicited an attention shift to a target determines *how* this target is processed.

More concretely, if previous experience of the meaningfulness of speaker gaze, in particular with respect to identifying intended referents (e.g., Griffin, 2001; Meyer, Sleiderink, & Levelt, 1998), causes such gaze cues to elicit inferences about referential intentions, then a certain order of gaze cues is predicted to elicit expectations for that same order in spoken references (Staudte & Crocker, 2011). That is, a congruent order of gaze cues and spoken references would facilitate listeners' comprehension and validation of such an utterance compared to neutral or incongruent stimuli (here *incongruent* is used in terms of the mismatched linear order of two verbal references and the concurrent gaze cues, henceforth called *reverse*). While there is obviously no such helpful cue in a neutral condition, a reverse sequence of gaze cues would effectively provide counter-predictive cues which may *not* be applied flexibly to the utterance if gaze is considered to reflect an immediate intention to refer to that object. Thus, a reverse sequence of gaze cues and verbal references may be counter-predictive and therefore informative and yet it could be disruptive to comprehension compared to a congruent cue or even no cue at all. In contrast, speaker gaze may be treated like other visual cues, such as arrows, which direct visual attention in a potentially very similar way but carry no such bias or requirement for a congruent order of cues as they do not lead to inferences of referential intentions. In this case, both cues would be exploited for comprehension in a maximally efficient manner – when the experimental design and task assigns a temporary benefit to them. This benefit is potentially available from congruent as well as reverse cue sequences such that listeners' comprehension may benefit almost equally.

2. Experiment 1

To investigate whether listeners infer referential intentions from speaker gaze, Experiment 1 examined whether an agent's gaze cues need to be sequentially aligned with corresponding referential speech cues (in the way human gaze typically precedes referring expressions) in order to be beneficial. In contrast, if gaze was a purely visual cue, a "misaligned" sequence of cues might also be beneficial since agent gaze still draws attention to mentioned objects in the scene, generally increasing those objects' salience.

We therefore manipulated sequential alignment of a combined *gaze and head movement*[1] with speech cues to assess the influence of this alignment on comprehension, as measured by response times for utterance validation. Specifically, the factor "Cue Order" had three levels: The sequence of two referential gaze cues and two referential nouns was either congruent (Fig. 1), reverse to each other, or neutral (gaze straight ahead, as in Fig. 1a; for sample stimuli see also Videos S1–S3 in the supporting information available on-line). Agent gaze was always directed to mentioned objects only, that is, gaze cues were invariably related to utterance content and therefore potentially informative and meaningful to the listener even if reverse to the verbal references. Agent gaze was, however, not required to determine utterance validity with respect to the scene.

2.1. Method

2.1.1. Participants

Twenty-four native speakers of German, all but one enrolled at Saarland University as students, took part in this study (mean age 26.8, 16 females). All participants reported normal or corrected-to-normal vision.

2.1.2. Materials

We created 1920 × 1080 resolution video-clips showing the virtual character *Amber* (Heloir & Kipp, 2009) located behind a table. Each video showed seven objects on the table which differed in shape and color. *Amber* performed a sequence of head and eye movements consecutively towards two objects that she also mentioned in a simultaneous utterance, e.g., "*Das Ei ist größ er als der Quader*" (English: "*The egg is taller than the box*"; see also Fig. 1). The utterance was a synthesized German sentence using the Mary TTS system (Schroeder & Trouvain, 2003).

Each item appeared in three conditions (congruent/reverse/neutral), so six lists were created, accounting for three within-subjects conditions and their counter-balanced versions. Each list contained 24 items and 36 fillers, of which 24 contained false utterances to motivate the validation task as items were always true. Each participant was assigned to one list, with the order of item trials randomized individually. That is, participants saw each item only in one condition.

2.1.3. Task and procedure

An EyeLink II head-mounted eye-tracker monitored participants' eye movements on a 24-inch monitor. Before the experiment, participants received written instructions explaining procedure and task: They were asked to attend to the presented videos and judge whether or not *Amber's* statements were valid scene descriptions. In order to provide a cover story for this task, participants were further told that the results were used as feedback in a machine learning procedure to improve the agent's performance. The entire experiment lasted approximately 25 min.

2.1.4. Analysis

Videos were segmented into labeled regions containing the objects referred to by the first noun ("*egg*") and the second noun ("*box*") and *Amber's* head. These regions were coded as circles around objects, containing the whole object but never overlapping with another object. This introduces some item variance since some objects are bet-

[1] Using such a combination of cues is only natural, as humans typically move their head *and* eyes as well. This does not conflict with the overall interpretation of gaze as speaker-inherent cue reflecting visual attention. Rather, gaze becomes more pronounced through head movements and can be followed peripherally (as e.g., in Hanna & Brennan, 2007).

(a) Agent gaze towards camera, (b) then to the EGG, (c) and to the BOX, before moving back up.

Fig. 1. Sample stimulus from Experiment 1. An utterance such as "*The egg is taller than the box.*" Was accompanied with either congruently ordered (as depicted), reversely ordered (straight ahead, then towards the BOX, the EGG, and back up) or neutral **gaze** cues.

ter captured by a circle than others. However, counterbalancing (in each scene there was always one larger and one smaller object mentioned, in both orders) as well as including random item effects should level out any potential variance and yield a meaningful measure for comparison between cue types.

We recorded participant fixations on these regions and report inspection probabilities, that is, whether a subject inspected a particular object or not in each of the following time windows: CUE1 stretched from the onset of the initial gaze cue to 100 ms after onset of the first noun ("*egg*") with a duration of 530 ms; N1 was adjacent to CUE1, starting at 100 ms after noun onset (referential eye movements are not to be expected until 100 ms after noun onset, Altmann, 2011) and with the duration of the respective noun (on average 386 ms); CUE2 began with the onset of the second gaze cue, 1030 ms prior to noun onset ("*box*"), and again lasted 530 ms; N2 was analogous to N1 and contained the second noun (on average 385 ms). Cueing intervals are depicted as shaded areas in Fig. 2. Further, the elapsed time between the second noun onset and the moment of the button press was considered as response time (RT). Accuracy was checked for differences between conditions before wrong button presses (13.6%) and outliers (3.4%) were removed from RT analysis. Inferential statistics were carried out using mixed-effects models from the lme4 package in R (Baayen, Davidson, & Bates, 2008). Specifically, we used logistic regression for modeling binary data such as accuracy counts and object inspections and linear regression for response times. Further, main effects and interactions were determined through model reduction which assesses the contribution of a predictor or interaction to a fitted model by running a χ^2-comparison between models with and without the particular predictor(s) (see, for instance, Jaeger, 2008).

2.2. Results and discussion

2.2.1. Button presses

False responses were relatively frequent in this study and distributed across conditions as follows: congruent = 16, neutral = 24, reverse = 35. A logistic regression model fitted to the data revealed a main effect of condition ($\chi^2(2) = 10.19,\ p < .01$) with reverse gaze eliciting significantly fewer correct responses than congruent trials (Coeff. = −1.02, SE = 0.33, Wald Z = −3.06, $p < 0.01$).

On average, listeners needed 1809 ms to respond to a neutral item, 1340 ms to respond to a congruent and 2062 ms to respond to a reverse item (mean response times are also depicted by dark gray bars in Fig. 4 further down). Using ANOVA to find the minimal model fitting the log-transformed RT data,[2] revealed a main effect of Cue Order on response times ($F = 95.64,\ df = 2$). The resulting model is shown in Table 1. It further revealed significant differences between the neutral and the congruent condition (Coeff. = −0.30, SE = 0.04, t = −7.52, $p < 0.001$), and also between the neutral and the reverse condition (Coeff. = 0.10, SE = 0.04, t = 2.51, $p < 0.05$; p-Values were calculated through Markov chain Monte Carlo (MCMC) sampling).

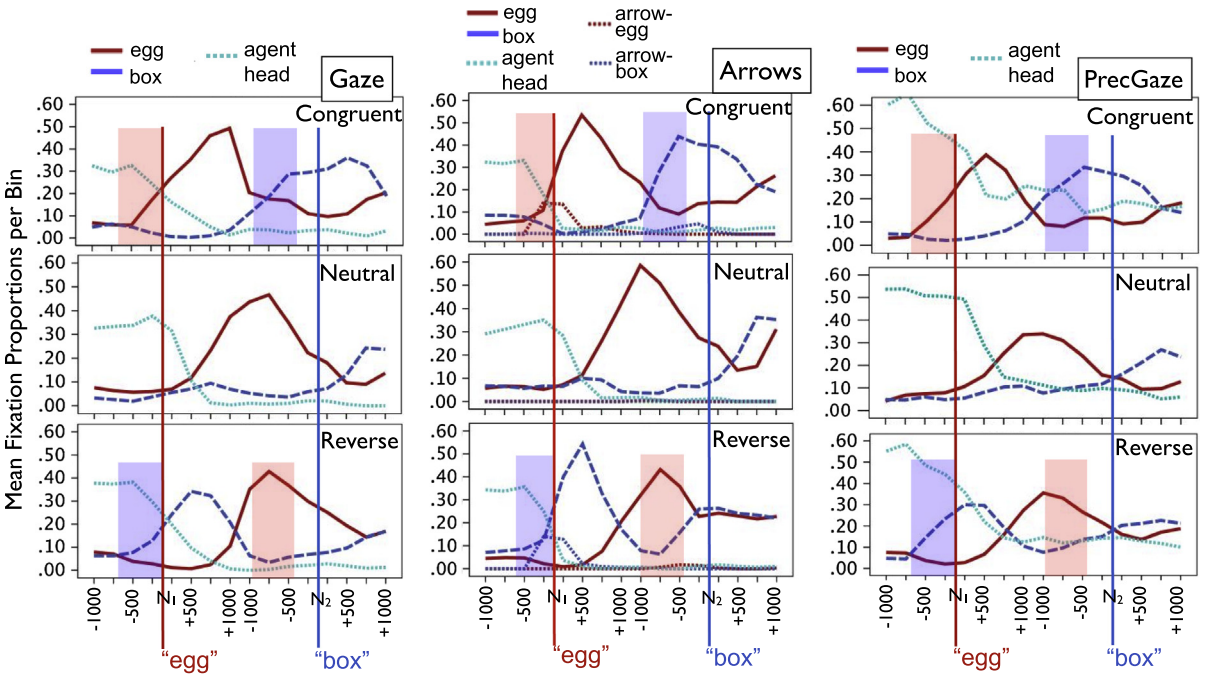
The overall means suggest that listeners could not use reversed gaze to facilitate sentence validation. However, since there could be development or potential adaptation of listeners to the experimental task and conditions masked by the overall mean, we included the experimental block in which a trial appeared (first or second half of the experiment) as a factor in our model. There was a main effect of Block (again revealed by model reduction: $\chi^2(1) = 4.26,\ p < .05$), showing that listeners became faster in general but, more importantly, there was no interaction of Block and Cue Order.

2.2.2. Eye movements

The time curves in Fig. 2a plot listeners' fixations towards the EGG, the BOX and *Amber's* head while *Amber* looks towards these objects and utters her description. In the top graph (congruent condition), *Amber's* first gaze movement towards the EGG (CUE1) is marked by the first shaded area prior to the mentioning of the noun "*egg*". The second gaze cue (CUE2) is marked by the second shaded area. This pattern is reversed in the reverse condition.

The plots clearly show that listeners followed *Amber's* gaze towards the corresponding objects and this interpretation is supported by inferential statistics using logistic regression for the inspection data. Already before the onset of the "*egg*" (CUE1), inspections on the EGG were significantly more likely in the congruent condition than in the reverse (Coeff. = −4.43, SE = 1.03, Wald Z = −4.29, $p < 0.001$) or neutral condition (Coeff. = −2.46,

[2] Generally, non-transformed data yielded similar main effects across all experiments.



(a) Lines plot fixations to *Amber's* head and to the EGG and the BOX as *Amber's* utterance *"The egg is taller than the box"* unfolds. The plots are aligned to the onset of "egg" and to the onset of "box"

(b) Depicted are fixations to *Amber's* head, the EGG and BOX, and to the regions containing the arrows, across the unfolding utterance.

(c) Because of the simplified scene and the lower distance between head and objects in Exp3, we cannot draw direct comparisons to the fixation patterns in Exps 1 and 2. The fixation graph is given here for convenience.

Fig. 2. Time curves from Experiments 1, 2a and 3.

Table 1
Final model fitted to response time data of Experiment 1 with a final set size of 455 data points. The last column shows *p*-Values calculated through Monte-Carlo-sampling.

Predictor	Coeff.	*SE*	*t*-Value	p_{MCMC}
(Intercept, neutral)	7.52	0.087	86.20	<.001
Order-congruent	−0.30	0.039	−7.52	<.001
Order-reverse	0.10	0.041	2.51	<.05
Block-second	−0.07	0.034	−2.06	<.05

$Model: log(RT) \sim CueOrder + Block + (1|subject) + (1|item)$.

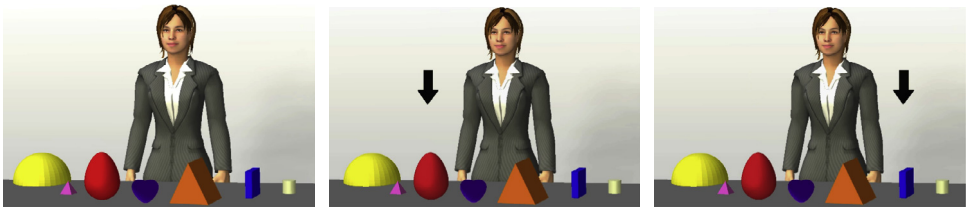
SE = 0.42, Wald Z = −5.84, $p < 0.001$). Similarly, the BOX was inspected more frequently in the reverse condition compared to congruent (Coeff. = −4.28, *SE* = 1.02, Wald Z = −4.22, $p < 0.001$) or neutral agent gaze (Coeff. = −2.46, *SE* = 0.45, Wald Z = −5.51, $p < 0.001$). In the neutral condition, in contrast, inspections on the EGG and BOX were equally likely during CUE1 (probability of 0.03 for both objects).

The same pattern was observed for the adjacent noun interval N1. Along with the time curve of the neutral condition, this suggests that interval N1 was still capturing gaze-following behavior and that referential eye movements (to the EGG) were largely delayed beyond noun offset. Crucially, we observed similar gaze- and speech-following patterns for the time windows CUE2 (and N2): as *Amber* looked at the BOX (congruent condition), for instance, inspections to the BOX became most likely while inspections to the EGG become least likely in this condition.

In summary, manipulating Cue Order created a mismatch between visual and spoken references which enabled us to observe which cue listeners followed initially and how they recovered from such a mismatch. The accuracy as well as the response time data suggest that people found the congruent condition easiest to process and the corresponding eye movements suggests that this was the case because listeners followed *Amber's* gaze and used it to anticipate the intended next referent. In the reverse condition, listeners were slowest and most often wrong which suggests that *Amber's* reversely ordered gaze cues disrupted the comprehension process. Even though the reverse condition technically provided information about both referents of the sentence earlier than the other two conditions (first agent gaze towards BOX, then mentioning of the "*egg*"), listeners were obviously unable to adapt and make use of this information to predict the mentioning of the box – and instead were persistently disrupted by the misaligned order of *Amber's* referential gaze and speech cues.

3. Experiment 2a

In this study, we replaced *Amber's* gaze cue with an arrow appearing above the corresponding object (see Fig. 3 and Videos S4–S6 in the supporting material on-line). This manipulation sought to reveal whether the facilitating and disruptive effects of gaze found in Experiment 1 were caused by the elicited visual attention shifts per se, or



(a) Agent gaze towards camera, (b) with arrow pointing to the EGG, (c) and then to the BOX, before disappearing.

Fig. 3. Sample stimulus from Experiment 2. The sample utterance "*The egg is taller than the box.*" Is accompanied with again either congruently ordered (as depicted), reversely ordered or neutral (i.e., without) **arrow** cues.

whether they were caused because listeners inferred the agent's intention to mention the fixated object from the gaze cue. Assuming that both gaze and arrows (reflexively) direct visual attention in a similar behavioral manner (Bayliss & Tipper, 2005; Ristic et al., 2002, 2007; Tipples, 2008), the Visual Account would predict identical effects on comprehension for arrow and gaze cues, whereas an intentional reading of gaze – but not arrows – would predict adaptation to the utility of the arrow cue regardless of order (as in the case of counter-predictive cues).

3.1. Method

3.1.1. Participants and procedure

Another twenty-four native speakers of German, of which all but one were students at Saarland University, took part in this study (mean age 24.6, 15 females). Again, all reported normal or corrected-to-normal vision. The presence of the arrows was explained to be a cue for *Amber*, showing her which objects she should talk about, and that sometimes she would not adhere to these cues. Task and Procedure were otherwise identical to Experiment 1.

3.1.2. Materials and analysis

The number and constitution of stimuli was identical to Experiment 1 except for the actual cue. That is, the gaze movement of *Amber* was replaced by an arrow above the respective object for the same onset and duration that gaze cues provided in Experiment 1. The neutral condition in Experiment 1 contained no directional cue and was identical for the current experiment. Consequently, spatial and time regions used in this experiment were identical to Experiment 1 but further included the two regions containing the arrows themselves. Again, trials with false responses (8.9%) and outliers (1.8%) were removed from RT analysis.

3.2. Results and discussion

3.2.1. Button presses

False responses were not so frequent in this study and distributed across conditions as follows: congruent = 9, neutral = 22, reverse = 18. A logistic regression model fitted to the data showed a main effect of condition ($\chi^2(2) = 6.76,\ p < .05$) with only neutral trials eliciting significantly fewer correct trials than the congruent condition (Coeff. = −1.01, *SE* = 0.43, Wald Z = −2.34, $p < 0.05$).

Mean response times of this experiment are displayed by Fig. 4 in direct comparison to the means from Experiment 1. Model reduction revealed, firstly, that including including random slopes for subject, but not item, provided a better fit of the model, and secondly, that Cue Order had a main effect on response times ($\chi^2(2) = 54.39,\ p < .001$). The model is shown in detail in Table 2. The pairwise comparisons among predictor levels revealed a significant difference between the neutral condition and both the congruent and the reverse conditions. This time, however, listeners were faster in both the congruent as well as the reverse condition (negative coefficients) compared to the neutral condition. Further, we further found a main effect of Block onto response times ($\chi^2(1) = 10.76,\ p < .01$) and also an interaction with Cue Order ($\chi^2(2) = 11.51,\ p < .01$). This interaction was carried by a significant speed up in the reverse condition: from a mean of 1477 ms in the first experimental block to 1123 ms in the second block ($t = -4.71,\ p < 0.001$). This suggests that listeners *learned* to exploit the counter-predictive utility of the arrow cues.

A combined analysis treating both experiments as a between-subject manipulation of Cue Type (gaze versus

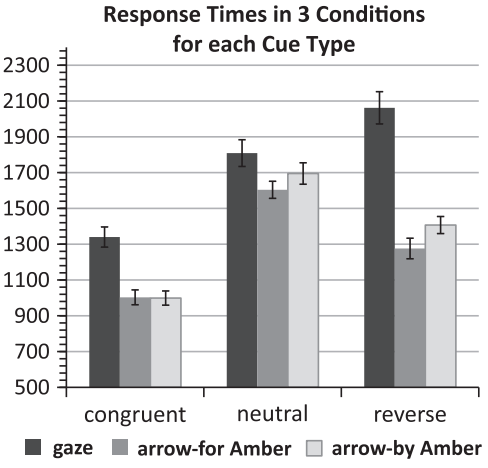


Fig. 4. Average response times in all three conditions, in direct comparison between the cue types *gaze* (Exp1), *arrow-for-Amber* (Exp2a) and *arrow-by-Amber* (Exp2b). Error bars reflect the standard error.

Table 2
Model fitted to response time data from Experiment 2a with a final set size of 491 data points. The last column shows *p*-values calculated through Monte-Carlo-sampling for the random intercept-only model.

Predictor	Coeff.	*SE*	*t*-Value	p_{MCMC}
(Intercept, neutral)	7.51	0.080	92.83	<.001
Order-congruent	−0.51	0.044	−11.62	<.001
Order-reverse	−0.27	0.066	−4.15	<.001
Block-second	−0.11	0.036	−3.29	<.01

$Model: log(RT) \sim CueOrder + Block + (1 + CueOrder|subject) + (1|item)$.

arrow) further revealed a main effect of Cue Type ($\chi^2(1) = 10.78,\ p < .01$) – listeners were generally faster in the arrow experiment – as well as an interaction between Cue Type and Cue Order ($\chi^2(2) = 39.08,\ p < .001$). The former results from the greater facilitation in congruent and reverse trials, neutral trials are similarly slow across Cue Type. The latter is due to the slowed response time for reverse gaze cues compared to speeded response time for reverse arrow cues.

3.2.2. Eye movements

The time curves plotted in Fig. 2b again show listener fixations on the EGG, BOX and *Amber's* head and, additionally, the arrow regions. Table 3 further shows the inspection data used for inferential statistics for both gaze and arrow cues in comparison. The probabilities that are critical for comparison are printed in bold-face. Importantly, inspection patterns were very similar to those in Experiment 1. Listeners again started a trial with fixating *Amber's* head even though it never moved. In fact, inspection probabilities for the object cued in CUE1 were extremely similar for both gaze and arrows. That is, when the EGG was gazed at (at this point no other information was provided yet), listeners inspected the EGG with a probability of 0.376. When an arrow pointed at the EGG, it was inspected with a probability of 0.317 (cf. interval CUE1, congruent condition, shown in row 2 in Table 3). Similarly, agent gaze to the BOX elicited an inspection probability of 0.359 and arrows resulted in an inspection probability of 0.343 (reverse condition, row 4).

Table 3
Inspection probabilities for the objects EGG and BOX, separately for each Experiment (Exp1/gaze and Exp2a/arrow) and in each condition, in time intervals CUE1, N1, CUE2 and N2. The probabilities reflect the proportion of trials in which at least one inspection to the respective object was performed by the participant.

Row		Object **egg**		Object **box**	
	CUE1	Gaze	Arrow	Gaze	Arrow
1	Congruent				
2	(cue to **egg**)	**0.376**	**0.317**	0.005	0.048
	Neutral				
3	(no cue)	0.049	0.059	0.032	0.076
	Reverse				
4	(cue to **box**)	0.005	0.022	**0.359**	**0.343**
5	"**N1 egg**"				
6	Congruent	**0.360**	**0.430**	0.011	0.016
7	Neutral	0.162	0.157	0.114	0.124
8	Reverse	**0.044**	**0.022**	0.354	0.403
9	**CUE2**				
	Congruent				
10	(cue to **box**)	0.167	0.091	**0.285**	**0.446**
	Neutral				
11	(no cue)	0.411	0.292	0.049	0.054
	Reverse				
12	(cue to **egg**)	**0.431**	**0.453**	0.083	0.088
13	"**N2 box**"				
14	Congruent	0.134	0.161	**0.322**	**0.151**
15	Neutral	0.070	0.092	0.162	0.216
16	Reverse	0.138	0.193	**0.111**	**0.166**

In fact, in the whole interval CUE1, we found only one significant effect (and none in N1) of Cue Type on eye movements: an interaction of Cue Type and Cue Order on the object BOX ($\chi^2(2) = 9.63,\ p < .01$). This interaction, however, was carried by differences between inspection probabilities on the object that was neither cued nor mentioned yet (BOX, in the congruent condition with $p = 0.057$ and the neutral condition with $p = 0.076$, which contains no cue at this stage and is identical in both experiments; details are shown in rows 2 and 3 respectively).

In the subsequent time intervals CUE2 and N2, some differences in inspection patterns to cued or mentioned objects could be observed: in interval CUE2, for instance, arrows in the congruent condition elicit more inspections to the cued BOX than gaze (Coeff. = −0.71, *SE* = 0.22, Wald Z = −3.24, $p < 0.001$). Possibly, the more gradual onset of the (sweeping) gaze cue – under the influence of previous cueing and concurrent speech – resulted in delayed identification and thus fewer inspections on the cued object.

Since the minor differences in the inspection data of Experiments 1 and 2a offer no apparent explanation of the substantial RT effects, we re-examined the fixation data plotted in Fig. 2: in the arrow experiment only, the graphs suggest anticipatory fixations back to the BOX *before* it was actually mentioned in N2 in the reverse condition. These early looks to the BOX prior to its mention suggest that listeners have understood that when the BOX was cued sentence-initially and had not been mentioned yet it is most likely mentioned next. Therefore, such anticipatory eye-moments index the ability to actively remap reverse arrow cues and they potentially facilitate comprehension of the upcoming reference. A correlation analysis of the actual time of the first fixation back to the BOX (from 1000 ms prior to "box" onset onwards) indeed revealed an $r = 0.498$ Pearson correlation ($p < 0.001$) between the first fixation time and the response time. That is, the earlier the first fixation back to the BOX occurred, the more likely was a short response time. This correlation held for both experiments individually as well as for a combined analysis. Crucially, however, the mean first fixation back to the BOX occurred significantly earlier in the arrow study (92 ms after noun onset, i.e., planned before noun onset) than in the gaze study (662 ms after noun onset, $p < 0.001$).

In short, the response time results of Experiments 1 and 2a show, firstly, that listeners were generally faster to determine the agent's utterance validity in Experiment 2a than in Experiment 1. This speed up, especially in the congruent condition, may be attributable to a lower cognitive load when integrating language with iconic directional and purely visual cues, such as the arrows employed in Experiments 2a, in contrast to integrating language and speaker

gaze. Secondly, and more importantly, listeners were able to use also reverse arrow cues to facilitate comprehension. In contrast to Experiment 1, where reverse gaze elicited wrong responses more often and significantly longer RTs than in the neutral and congruent gaze conditions, reverse arrows led to speeded responses compared to the neutral condition. Moreover, this patterns was enhanced in the second half of Experiment 2a, indicating a learning effect for improved use of reverse cues which was absent in Experiment 1.

While the eye movement results show largely a similar pattern of visual attention shifts across both Experiments, reverse arrows elicited listener fixations to the object mentioned second earlier than reverse gaze. This provides further support for our claim that listeners were able to exploit reverse arrow cues to predict the second referent and respond faster compared to the neutral condition.

In sum, the findings suggest a clear difference in how listeners exploit gaze versus arrow cues, as manifest in the reverse order condition, but there are other possible explanations for these findings. Firstly the difference may be due to the association of intentions with the cue. While gaze was clearly perceived as a speaker inherent cue, the arrow cues were explained to be a cue *for* the agent speaker, provided *by* the experimenter. Thus, arrows offered a speaker-unrelated, potentially non-intentional cue which may be one reason for the systematic difference in how listeners exploited gaze and arrows. Secondly, despite the obviously similar patterns of visual attention shifts, the arrow cue appears closer to each object and potentially selects an object more saliently and more clearly than does the agent gaze. Even though sentence-initial fixation and inspection patterns indicate that there is no difference between the ability to follow each cue type, the (learned) mapping of cues and anticipation for the target object may still relate to cue precision. To address these two possible explanations, we conducted two further experiments. Experiment 2b tackles the question whether RT differences originate from different intentions associated with the cues; and Experiment 3 addresses the potential precision issue by presenting a simplified gaze study which is designed to better match the precision of the arrow baseline.

4. Experiment 2b

To investigate whether the effects of Cue Type on response times as described above may rather have been carried by the possibly different motivations to follow arrows and gaze cues (gaze was produced by the agent speaker and possibly perceived as more error-prone than arrows), we conducted a third experiment. In this study, we replicated Experiment 2a with an alternative instruction: arrows were now explained to be a means for *Amber* to display her current interest in an object. That is, we tested whether inviting a more speaker-related, intentional reading of arrows would lead also to a more "intentional" use by listeners, as in the case of speaker gaze (Experiment 1), or whether listeners would still exploit these arrows flexibly, as in Experiment 2a.

4.1. Method

4.1.1. Participants and procedure

A third set of twenty-four native speakers of German took part in this study (mean age 25.33, 20 females). Again, all were students at the Saarland University and reported normal or corrected-to-normal vision. The presence of the arrows was explained to be a cue by *Amber*, signaling the participant which object she was currently interested in.

4.1.2. Materials and analysis

The number and constitution of stimuli was identical to Experiment 2a. Again, trials with false responses (11.27%) and outliers (2.9%) were removed from RT analysis.

4.2. Results and discussion

4.2.1. Button presses

As in Experiment 2a, false responses were not so frequent and distributed across conditions as follows: congruent = 13, neutral = 21, reverse = 28. Again, a logistic regression model fitted to the data showed a main effect of condition ($\chi^2(2) = 6.96,\ p < .05$), with congruent trials eliciting significantly more correct/fewer incorrect responses than reverse trials (Coeff. = 0.91, *SE* = 0.36, Wald Z = 2.52, $p < 0.05$).

Importantly, response times similar to Experiment 2a were obtained, with mean values of 999 ms, 1695 ms and 1407 ms in the congruent, neutral and reverse condition, respectively (also shown also in Fig. 4 and in Table 4), as well as a main effect of Cue Order ($\chi^2(2) = 39.28,\ p < .001$) and Block ($\chi^2(2) = 4.92,\ p < .05$; though no interaction with Block). In a combined analysis of Experiments 2a and 2b including Cue Order and Cue Type (arrow for *Amber* versus arrow by *Amber*), we found the main effect of Cue Order ($\chi^2(2) = 265.32,\ p < .001$) but no main effect of, or interaction with, Cue Type ($\chi^2(1) = 0.09,\ p = .759$ and $\chi^2(2) = 3.09,\ p = .21$, respectively). In contrast, the comparison of response times in Experiment 1 and 2b reveals similar effects (main effect of Cue Type, $\chi^2(1) = 7.14,\ p < .01$, and interaction with Cue Order, $\chi^2(2) = 37.08,\ p < .001$) as the comparison between Experiments 1 and 2a above.

4.2.2. Eye movements

Eye movements were practically identical to those in Experiment 2a and are therefore omitted here. Graphs can be found in the supporting online material (Fig. S7).

Table 4
Final model fitted to response time data of Experiment 2b with a final set size of 477 data points. The last column shows *p*-Values calculated through Monte-Carlo-sampling for the random intercept-only model.

Predictor	Coeff.	*SE*	*t*-Value	p_{MCMC}
(Intercept, neutral)	7.45	0.055	134.210	<.001
Order-congruent	−0.61	0.070	−8.71	<.001
Order-reverse	−0.27	0.087	−3.10	<.05
Block-second	−0.09	0.038	−2.38	<.05

$Model: log(RT) \sim CueOrder + Block + (1 + CueOrder|subject) + (1|item)$.

The lack of a main effect or interaction with CueType on response times and the similarity of the eye movements across Experiments 2a and 2b suggests that introducing arrow cues as a cue generated by the experimenter (Experiment 2a) or by the speaker (Experiment 2b) did not significantly influence how listeners exploited these arrow cues. That is, despite offering different (non-)intentional readings of the arrow cue, listeners treated it identically across Experiments 2a and 2b.

5. Experiment 3

The design of the arrow cues necessitated a change in cue position relative to the referents, which may have resulted also in a change in cue precision (compare Figs. 1 and 3). Experiment 3 was therefore designed to employ and test a gaze cue whose precision better matched that of the arrow cues. The higher precision was achieved (a) by simplifying the visual scene, i.e., reducing the number of objects and by introducing more space between them, and (b) by decreasing the space between the table top and the agent head such that its gaze and head movements were more pronounced and more distinguishable. Due to the changes in the visual scene, the results cannot be compared directly to Experiments 1 and 2, as for instance in combined statistical analyses, but the general patterns are expected to reveal whether or not participants are now able to (learn to) remap gaze in reverse sequences in order to anticipate the second referent, as is the case in the arrow studies.

Because cue precision was an important feature in this study, we conducted a pre-test of the stimuli. In this pre-test, we showed the item videos without audio to participants and asked them to write down which two objects had been looked at by the agent. Accuracy was 91.7%, which may still underestimate the actual accuracy in the experiment where participants got immediate verbal feedback and potentially learned to better read the gaze gestures. We then also conducted a post-test of the stimuli used in Experiment 2a and 2b (arrows) in order to test wether cue precision was indeed similarly high in both studies. The accuracy for arrow stimuli was 98.4%.

5.1. Method

5.1.1. Participants and procedure

Another twenty-four native speakers of German, all of which were students at Saarland University, took part in this study (mean age 23.8, 13 females). Again, all reported normal or corrected-to-normal vision. Task and Procedure were identical to Experiment 1.

5.1.2. Materials and analysis

The number and constitution of stimuli was similar to Experiment 1 except for the decreased number of distractors displayed in each scene. Specifically, only four objects instead of previously seven were located on the table. Across items, objects in all four positions were referred to, but always such that one referent was on the left and the other referent on the right side of the table. The two other objects were each a potential size competitor for one of the referents (see Fig. 5 for a sample scene). This was to ensure that participants had to wait and hear the head noun before being able to validate the statement (if no other object, other than the BOX, was shorter than the EGG, then participants would not have to wait to hear "egg"). Trials with false responses (17.1%) and outliers (1.9%) were removed from RT analysis.

5.2. Results and discussion

5.2.1. Button presses

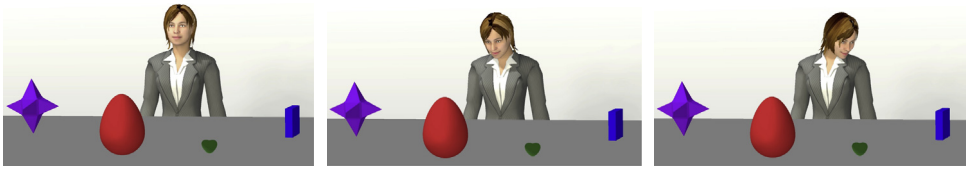
False responses were relatively frequent in this study and distributed across conditions as follows: congruent = 25, neutral = 27, reverse = 46. A logistic regression model fitted to the data revealed a main effect of condition ($\chi^2(2) = 15.41,\ p < .001$) with reverse gaze eliciting significantly fewer correct/more false responses than both congruent (Coeff. = 1.26, $SE = 0.36$, Wald $Z = 3.41,\ p < 0.001$) and neutral trials (Coeff. = 1.10, $SE = 0.35$, Wald $Z = 3.09,\ p < 0.01$). In sum, accuracy in both gaze studies (Experiments 1 and 3) is lower than in the arrow study (Experiment 2a). That is, a combined analysis reveals a main effect of CueType with significantly fewer correct responses for gaze, Coeff. = −0.52, $SE = 0.27$, Wald $Z = -1.92$, $p = 0.054$, and precise-gaze, Coeff. = −0.85.26, $SE = 0.26$, Wald $Z = -3.24$, $p < 0.01$, compared to arrows.

Mean response times of this experiment are displayed by Fig. 6 in direct comparison to the means from Experiment 2a. Model reduction revealed, firstly, that including random slopes for subject, but not item, provided a better fit of the model, and secondly, that Cue Order had a main effect on response times ($\chi^2(2) = 19.16,\ p < .01$). The model is shown in detail in Table 5. The pairwise comparisons among predictor levels revealed again a significant difference between the neutral condition and the congruent condition and, just as in Experiment 2, between the neutral and the reverse condition. Congruent and reverse gaze trials did not differ significantly.

As in Experiment 2a (arrows), we also found a main effect of Block onto response times ($\chi^2(1) = 5.63,\ p < .05$) and an marginally significant interaction with Cue Order ($\chi^2(2) = 5.05,\ p = .080$). This interaction was carried by a significant speed up in the reverse condition: From a mean of 1499 ms in the first experimental block to 1365 ms in the second block ($t = -2.33,\ p < 0.05$). Even though the speed up is not as large as in the arrow study, this finding suggests that listeners also *learned* to better exploit counter-predictive gaze cues.

5.2.2. Eye movements

The time curves plotted in Fig. 2c show listener fixations on the EGG, BOX and *Amber's* head and, additionally, the arrow regions. Fixation pattern look very similar to those in Experiments 1 and 2 but because of the simplified scene and the lower distance between head and objects, we cannot draw direct comparisons to the inspection frequencies in the previous experiments. Instead, we examined whether participants also showed the kind of anticipatory eye-movement behavior that we found in the arrow study (Experiment 2a). In particular, we also



(a) Agent gaze towards camera, (b) to the EGG, (c) then to the BOX, before moving back up.

Fig. 5. Sample stimulus from Experiment 3. The sample utterance "*The egg is taller than the box.*" With gaze (or not) towards the EGG and the BOX. Only two other objects are displayed in this simplified scene.

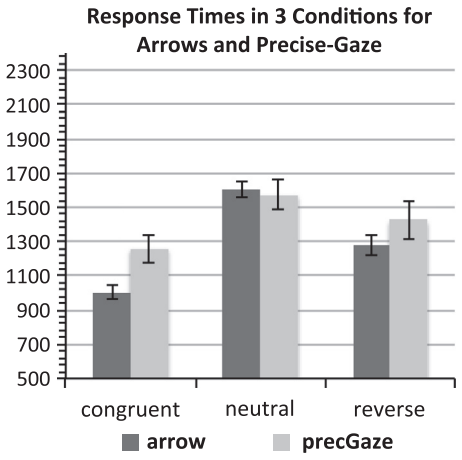


Fig. 6. Average response times in all three conditions, in direct comparison between the cue types *arrow* (Exp2a) and *precise-gaze* (Exp3). Error bars reflect the standard error.

Table 5
Model fitted to response time data from Experiment 3 with a final set size of 463 data points. The last column shows *p*-Values calculated through Monte-Carlo-sampling for the random intercept-only model.

Predictor	Coeff.	*SE*	*t*-Value	p_{MCMC}
(Intercept, neutral)	7.375	0.100	73.66	<.001
Order-congruent	−0.215	0.049	−4.32	<.001
Order-reverse	−0.104	0.078	−1.34	<.05
Block-second	−0.094	0.039	−2.40	<.05

$Model: log(RT) \sim CueOrder + Block + (1 + CueOrder|subject) + (1|item)$.

extracted the time of the first fixation back to the target referent (again in a time window starting 1000 ms prior to N2, i.e., the first fixation to the BOX from 1000 ms before onset of "box" onwards) in the reverse condition. In Experiment 2a we found a mean first fixation time of 92 ms after "box" onset suggesting that on average participants programmed a fixation to the referent before it was mentioned. In Experiment 1, the mean first fixation occurred at 662 ms after "box" onset, so significantly later. In the current study showing precise gaze cues, the mean first fixation occurred at 187 ms after "box" onset which is marginally significantly faster than in the original gaze experiment (Experiment 1, $p = 0.08$) and no different from the arrow study. As in each of the previous experiments, this (anticipatory) first fixation time correlates with response times ($r = 0.403$ Pearson correlation ($p < 0.01$)). For the sake of completeness, Table 6 provides an additional overview of these first fixations and their average timing for each experiment.

Table 6
Average time of first fixation back to the BOX with SEs for all experiments.

Experiment	CueType	FixTime (ms)	*SE*
Exp1	Gaze	662	157
Exp2a	Arrow-for-Amber	92	109
Exp2b	Arrow-by-Amber	341	380
Exp3	Precise-gaze	187	138

6. General discussion

In four experiments, we investigated whether the benefit of speaker gaze for utterance comprehension was due to the referential intention that listeners inferred from the gaze cue, or whether the induced shift of the listener's visual attention *alone* yielded this benefit. To tease apart these two potential accounts of gaze utility, we compared the effects of speaker gaze on listeners' visual attention shifts (inspection patterns) and their utterance comprehension (response times) with the effects of a visual baseline cue, namely arrows.

Experiments 1 and 2 revealed that reverse gaze disrupted comprehension while reverse arrows facilitated comprehension, as revealed by reaction times to the sentence verification task. Crucially, this was the case despite both cues eliciting nearly identical gaze patterns. Indeed, the only difference between the two studies was the presence of an anticipatory gaze to the referent of N2 in the reverse arrow condition only, which we argue revealed the ability of listeners to "remap" the cue order for arrows, but not gaze. Experiment 2b then addressed the concern that differences in Experiments 1 and 2a might be due to the absence of a referential status for the arrow cue, but found no evidence to support this explanation.

In contrast, Experiment 3 considered whether visual precision of the two cues might be responsible for the observed difference. Since gaze cues (Experiment 1) were perceived as somewhat less precise in uniquely identifying the cued object[3] than arrows (Experiment 2a), this may

[3] Such visual imprecision may actually be considered a natural feature of gaze since eye balls are small and the distance to the viewed object is often at least as large as in Experiment 1. However, since speaker gaze and utterance are typically closely synchronized (i.e., congruent) such that listeners get rapid confirmation for their gaze-based predictions, imprecision is unlikely to have the same disruptive influence in natural conversations.

have negatively influenced the ability to remap the visual cue in the reverse gaze condition and exploit the fact that the object initially cued is the last one mentioned. Using new stimuli with increased gaze precision, Experiment 3 revealed that listeners were indeed able to remap gaze cues, as indicated by facilitated response times for both congruent and reverse gaze compared to neutral stimuli.

Some minor differences between the two cues in Experiments 2a and 3 were also detected: accuracy in both gaze studies (Experiments 1 and 3) is lower than in the arrow study (Experiment 2a). Moreover, the average first fixation to the object to be mentioned in N2 in the reverse condition occurred numerically later in Experiment 3 compared to Experiment 2a (187 ms versus 97 ms after noun onset), despite the simplification of the scene/task. The same holds for the learning effect (speed up from first to second block), which is smaller for precise-gaze, and the mean RT in the second block, which is slower for precise-gaze than for arrow cues. These minor variations may be attributed to the slight deviations in cue precision (91% for gaze versus 98% for arrows) and/or the scene design, or they may also hint at other, subtle differences in the mental processes and representations elicited by the two types of cues that are not revealed by the gaze and response time measures in our setting.

In sum, however, our findings suggest that when gaze is made as precise as arrows, there is *no qualitative difference* in their influence on comprehension. Specifically, this set of experiments was aimed to reveal whether the influence of speaker gaze on utterance comprehension goes beyond visual cueing (Intentional Account) or not (Visual Account). Under the Intentional Account, it was assumed that congruent gaze order alone would reflect speaker intentions, in contrast to arrows, and that reverse gaze order would therefore not facilitate – or even disrupt – comprehension. The fact that listeners can indeed remap gaze, like arrow cues, in the reverse condition and benefit from that in terms of speeded RTs provides strong support for the Visual Account, and no support for the Intentional Account. This result has further implications for other studies showing a beneficial effect of object-oriented speaker gaze (e.g., Hanna & Brennan, 2007; Nappa et al., 2009) in so far as such effects need not be unique to the presented (human) gaze cues and may in fact be replicated by employing any other visual cue with similar timing.

Acknowledgments

The research reported of in this paper was supported by the "Multimodal Computing and Interaction" Cluster of Excellence at Saarland University. Many thanks also to two anonymous reviewers for their greatly appreciated contributions.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2014.06.003.

References

Altmann, G. T. (2011). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psychologica, 137*(2), 190–200.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412.

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

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

Becchio, C., Bertone, C., & Castiello, U. (2008). How the gaze of others influences object processing. *Trends in Cognitive Science, 12*(7), 254–258.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition, 6*(5), 509–540.

Emery, N. (2000). The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews, 24*(6), 581–604.

Flom, R., Lee, K., & Muir, D. (Eds.). (2007). *Gaze-following: Its development and significance*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Friesen, C., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review, 5*(3), 490–495.

Friesen, C., Ristic, J., & Kingstone, A. (2004). Attentional effects of counterpredictive gaze and arrow cues. *Journal of Experimental Psychology: Human Perception and Performance, 30*(2), 319–329.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition, 82*(1), B1–B14.

Hanna, J., & Brennan, S. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language, 57*(4), 596–615.

Heloir, A., & Kipp, M. (2009). EMBR – A realtime animation engine for interactive embodied agents. In *Proc. of the 9th int'l conference on intelligent virtual agents*. Springer.

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

Langton, S. R., & Bruce, V. (1999). Reflexive visual orienting in response to the social attention of others. *Visual Cognition, 6*(5), 541–567.

Meltzoff, A. N., Brooks, R., Shon, A. P., & Rao, R. P. (2010). Social robots are psychological agents for infants: A test of gaze following. *Neural Networks, 23*(8–9), 966–972 (Social Cognition: From Babies to Robots).

Meyer, A., Sleiderink, A., & Levelt, W. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition, 66*(2), B25–B33.

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

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*(1), 3–25.

Ristic, J., Friesen, C. K., & Kingstone, A. (2002). Are eyes special? It depends on how you look at it. *Psychonomic Bulletin & Review, 9*(3), 507–513.

Ristic, J., Wright, A., & Kingstone, A. (2007). Attentional control and reflexive orienting to gaze and arrow cues. *Psychonomic Bulletin & Review, 14*(5), 964–969.

Schroeder, M., & Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology, 6*(4), 365–377.

Staudte, M., & Crocker, M. W. (2010). When robot gaze helps human listeners: Attentional versus intentional account. In S. Ohlsson & R. Catrambone (Eds.), *Proc. of the 32nd annual conf. of the cognitive science society*. Portland, OR: Cognitive Science Society.

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

Tipples, J. (2008). Orienting to counterpredictive gaze and arrow cues. *Perception & Psychophysics, 70*(1), 77–87.

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science, 10*(1), 121–125.

Trueswell, J., Lin, Y., Cartmill, E., Amstrong, B., Goldin-Meadow, S., Gleitman, L. (unpublished). *When timing is (almost) everything: Referential dynamics in parent-child interactions*. (Presented at CUNY 2013).

Vaidya, C. J., Foss-Feig, J., Shook, D., Kaplan, L., Kenworthy, L., & Gaillard, W. D. (2011). Controlling attention to gaze and arrows in childhood: An fMRI study of typical development and autism spectrum disorders. *Developmental Science, 14*(4), 911–924.