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

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