Acta Psychologica xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Acta Psychologica

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

Distinct effects of different visual cues on sentence comprehension and later recall: The case of speaker gaze versus depicted actions

Helene Kreysa[a,*,1], Eva M. Nunnemann[b,2], Pia Knoeferle[c,d,e,3]

[a] Institute of Psychology, Friedrich Schiller University Jena, Jena, Germany
[b] Cognitive Interaction Technology Center of Excellence, Bielefeld University, Bielefeld, Germany
[c] Department of German Studies and Linguistics, Humboldt-Universität zu Berlin, Germany
[d] Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Germany
[e] Einstein Center for Neurosciences Berlin, Germany

ARTICLE INFO

PsycInfo classification code:
2300 (Human Experimental Psychology)
Keywords:
Situated language processing
Spoken sentence comprehension
Anticipatory eye movements
Visual context
Gaze cueing
Visual-world paradigm

ABSTRACT

Language-processing accounts are beginning to accommodate different visual context effects, but they remain underspecified regarding differences between cues, both during sentence comprehension and subsequent recall. We monitored participants' eye movements to mentioned characters while they listened to transitive sentences. We varied whether speaker gaze, a depicted action, neither, or both of these visual cues were available, as well as whether both cues were deictic (Experiment 1) or only speaker gaze (Experiment 2). Speaker gaze affected eye movements during comprehension similarly early to a single deictic action depiction, but significantly earlier than non-deictic action depictions; conversely, depicted actions but not speaker gaze positively affected later recall of sentence content. Thus, cue type and cue-language relations must be accommodated in characterising real-time situated language comprehension and subsequent recall of sentence content.

1. Introduction

When speakers talk about their environment, they show a strong tendency to inspect objects before mentioning them (e.g., Bock, Irwin, Davidson, & Levelt, 2003; Griffin, 2004; Griffin & Bock, 2000; Kuchinsky, Bock, & Irwin, 2011; Meyer & Lethaus, 2004; Meyer, Roelofs, & Levelt, 2003). This referential nature of speech-related eye movements means that a speaker's gaze can help listeners to visually anticipate upcoming speech content, as has been shown in a number of different settings (e.g., Hanna & Brennan, 2007; Knoeferle & Kreysa, 2012; MacDonald & Tatler, 2013; Staudte & Crocker, 2011; Staudte, Crocker, Heloir, & Kipp, 2014). For instance, in a real-world eyetracking study by MacDonald and Tatler (2013) an experimenter instructed each participant on how to complete a block construction task. If the spoken instructions were ambiguous with regard to the referred-to block, participants were more accurate at selecting the correct block when the experimenter looked at it than when he did not. In fact, they actively sought the experimenter's gaze in the ambiguous language condition (when gaze cues were provided), but not with unambiguous instructions (cf. Nappa, Wessel, McEldoon, Gleitman, & Trueswell, 2009).

However, speaker gaze is only one of many cues that can assist comprehenders in assigning reference during spoken comprehension. Indeed, recent years have seen numerous studies on how a variety of individual cues affect the unfolding interpretation of spoken sentences (only a brief selection can be mentioned here; for reviews see Altmann, 2011; Huettig, Rommers, & Meyer, 2011; for theoretical accounts see Altmann & Kamide, 2009; Knoeferle & Crocker, 2006, 2007). Among these cues are linguistic ones, such as frequencies of acoustic and syntactic features (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Snedeker & Trueswell, 2004) and verb-argument relations (e.g., Altmann & Kamide, 1999), but also extralinguistic cues such as action affordances (Chambers, Tanenhaus, & Magnuson, 2004), arrows (Staudte et al., 2014), and visible actions (Knoeferle, Crocker, Scheepers, & Pickering, 2005). Most of these studies have made use of the so-called “visual world” paradigm, whereby participants' attention to an array of potential referents presented in front of them (e.g., Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) or on a screen (e.g., Altmann & Kamide, 1999) is interpreted as reflecting their unfolding linguistic interpretation of a simultaneously presented sentence. In this context, so-called “anticipatory” or “predictive” eye movements

* Corresponding author.
E-mail addresses: helene.kreysa@uni-jena.de (H. Kreysa), enunnemann@techfak.uni-bielefeld.de (E.M. Nunnemann), pia.knoeferle@hu-berlin.de (P. Knoeferle).
[1] Postal address: General Psychology and Cognitive Neuroscience, Institut für Psychologie, Friedrich-Schiller-Universität Jena, Am Steiger 3/Haus 1, D-07743 Jena, Germany.
[2] Postal address: CITEC, Inspiration 1, D-33615 Bielefeld, Germany.
[3] Postal address: Unter den Linden 6, D-10099 Berlin, Germany.

https://doi.org/10.1016/j.actpsy.2018.05.001
Received 23 February 2017; Received in revised form 15 January 2018; Accepted 5 May 2018

have played an important role, since they imply that some aspect of the situation allows comprehenders to identify upcoming referents before the speaker explicitly mentions them (for an early discussion of anticipatory fixation behaviour during language processing see Cooper, 1974). In the study by Knoeferle et al. (2005) for instance, participants rapidly integrated the identities of characters performing different actions with the meaning of the unfolding verb. This enabled them to work out correctly who was doing what to whom, despite temporary grammatical ambiguity of the presented sentence.

Such anticipatory eye movements have generally been interpreted as implying faster and more efficient comprehension, against the theoretical backdrop of a growing interest in the role of prediction for language processing (e.g., Altmann & Mirković, 2009; Chang, Dell, & Bock, 2006; Crocker, Knoeferle, & Mayberry, 2010; Federmeier, 2007; Huettig, 2015; Levy, 2008; Pickering & Garrod, 2013) and in fact as a core mechanism of cognitive processing in general (e.g., Clark, 2013; Friston, 2010). This interpretation rests on several assumptions. First, it assumes that speeded listener fixations to a target object (e.g., when cued by the speaker's gaze) imply that the reference to this object is built into the unfolding sentence representation, as originally described by Cooper (1974): “The existence of an anticipatory characteristic implies that people undergo an active online process of constructing hypotheses regarding the next successive informative item of speech, and then use the visual-motor system to test out those hypotheses prior to confirmation” (p. 104). An interesting observation in support of this assumption comes from Staudte and Crocker's (2011) Experiment 2. Participants in this study were asked to provide verbal corrections when the speaker (in this case a robot) produced a false description of the visual scene. In some cases, they corrected overtly true utterances which had been accompanied by incongruent speaker gaze, suggesting that they assumed the gaze behaviour reflected the speaker's underlying referential intention.

Second, there is the important question of whether anticipatory fixations by a listener actually facilitate task performance, i.e. whether they lead to more accurate responses or shorter response latencies if participants are required to respond overtly to the sentence. The literature to date is unclear on this point: While a few studies have shown such facilitation (e.g., Cooper, 1974; Staudte et al., 2014; Zhang & Knoeferle, 2012), others have reported either null-effects or a different effect pattern (e.g., Carminati & Knoeferle, 2013; Knoeferle & Kreysa, 2012). With regard to effects of eye gaze following, Hanna and Brennan (2007) report that participants in their study were in fact slower to reach for cued objects than in other eyetracking experiments, and even speculate that “this might reflect a cost of monitoring a partner's eye gaze” (p. 613).

Finally, there are the two related questions of whether all cues are equal and whether more cues are better for generating anticipatory fixations on the part of the listener. Although studies such as the ones mentioned above have shown that many different kinds of cue can rapidly inform comprehension, influential models of sentence processing have focused primarily on linguistic cues and have exploited just one of their properties, the probability with which a cue supports one interpretation over another (e.g., McRae, Spivey-Knowlton, & Tanenhaus, 1998). Potential differences in cue effects depending on the type of cue (such as speaker gaze or action depictions) have largely been ignored by modelling efforts to date. Situation model theories, for instance, offer a rich model of diverse aspects of our visual context, but are underspecified regarding potential differences in the type of a cue, and accordingly their effects on comprehension (Zwaan & Radvansky, 1998). The same is true for current processing accounts and frameworks of situated language processing (e.g., Altmann & Mirković, 2009; Huettig, Olivers, & Hartsuiker, 2011; Knoeferle & Crocker, 2006, 2007; Kukona & Tabor, 2011; but see Knoeferle, Urbach, & Kutas, 2014).

In fact, if cue type really does not matter (cf. Staudte et al., 2014; but see Knoeferle et al., 2014), then different cues, such as speaker gaze and action depictions, should influence comprehension to the same extent and in similar ways (assuming that they have a similar within-experiment frequency, precision, and come into play at the same point in time). Indeed, as described above, action depictions and speaker gaze both appear to be strong cues; each of them has individually been shown to outweigh other cues, albeit in different tasks. Depicted actions were prioritised over stereotypical knowledge in guiding comprehenders' visual anticipation during sentence listening (Knoeferle & Crocker, 2006). Similarly, although Staudte et al. (2014) did not find preferential processing of a virtual agent's gaze over simple arrow cues, Neider, Chen, Dickinson, Brennan, and Zelinsky (2010) showed that a gaze cue (a cursor overlaid on a scene, symbolising the gaze of a communicative partner) led to faster collaborative spatial search than verbal input.

Yet depicted actions and speaker gaze also differ in potentially important ways: Dynamic gaze shifts by a speaker can be processed reflexively (e.g., Teufel, Alexis, Clayton, & Davis, 2010) and peripherally (at least when they are accompanied by head movements, cf. Burton, Bindemann, Langton, Schweinberger, & Jenkins, 2009), and they are socially relevant as indicators of the speaker's referential intentions (Staudte et al., 2014). Additionally, gaze relates to the world in a different way than action depictions: Similar to several other visual cues such as arrows and movement or sudden onset of an object on the screen, gaze cues are deictic in the sense that they point directly to the upcoming referent (assuming that the listener follows the speaker's gaze). This deictic property may make them easier to process than cues more tightly relating to the linguistic content. For instance, benefitting from depicted actions as cues to upcoming sentence content involves a series of processing steps including recognising a particular action, linking it semantically to the verb, recognising the agent performing it, identifying suitable patients in the visual scene and integrating all of these identities into the representation of the unfolding sentence. The relationship between a depicted action and the verb of the sentence thus taps into the linguistic event representation of this verb, potentially activating its argument structure and thematic roles. To the extent that any of the mentioned differences in world-language relations between the two types of cue modulate their respective influence, we should see differential effects on unfolding comprehension and possibly also on subsequent recall of sentence content. Alternatively, finding no clear differences in the effects of these two cues could indicate that they contribute in a similar fashion to the unfolding interpretation and memory representation (see Staudte et al., 2014).

In addition to comparing different types of information affecting referential comprehension and investigating how two or more of such cues might interact, a secondary question raised in this paper was the subsequent recall of sentence content contingent on the availability of predictive cues during initial comprehension. This approach is relatively novel and must remain largely exploratory, but it follows from the previously described widely-held view that prediction is beneficial for comprehension processes. If this is indeed the case, one might expect that comprehension is not just speeded, but also that it is in some way “better”. For example, if a predictive cue allows early disambiguation between potential referents, the comprehender can spend relatively more time fixating and – presumably – processing the correct referent, and less time on the irrelevant one. This additional time could lead to better encoding, and subsequently to improved recall rates.

Against this theoretical backdrop, we compared the respective influences of two contextual cues on visual attention during spoken language comprehension and ensuing recall. In two studies, we pitted depicted actions against speaker gaze: Following an established paradigm (Knoeferle & Kreysa, 2012), we recorded participants' fixations as they watched videos of a speaker producing a transitive sentence about two virtual characters. Critically, we varied (a) whether the speaker shifted her gaze between the sentence referents, and (b) whether objects semantically related to the verb appeared between them. We reasoned that differences in the effects of the two cues could manifest behaviourally in the extent to which attention is allocated to the two