Effects of speaker gaze versus depicted actions on visual attention during sentence comprehension

Helene Kreysa (helene.kreysa@uni-jena.de)
Department of General Psychology, Friedrich Schiller University of Jena, Germany

Pia Knoeferle (knoeferl@cit-ec.uni-bielefeld.de)
Eva M. Nunnemann (eva@nunnemann.com)
Cognitive Interaction Technology Excellence Cluster
Department of Linguistics
Bielefeld University, Inspiration I, 33615, Bielefeld, Germany

Abstract

An eye-tracking study compared the effects of actions (depicted as tools between on-screen characters) with those of a speaker's gaze and head shift between the same two characters. In previous research, each of these cues has rapidly influenced language comprehension on its own, but few studies have directly compared these two cues or, more generally, distinct non-linguistic cues in their effects on real-time sentence comprehension. We investigated how participants used action tools and speaker gaze separately and in combination for visually anticipating the upcoming mention of a sentence referent. We discuss implications for accounts of visually situated language comprehension.

Keywords: eye tracking, spoken language comprehension, speaker gaze, depicted actions

Introduction

Recent years have seen numerous studies on how individual contextual cues affect the unfolding interpretation of spoken sentences (for recent reviews see Altmann, 2011; Huettig, Rommers, & Meyer, 2011; for theoretical accounts and computational models see Altmann & Kamide, 2009; Crocker, Knoeferle & Mayberry, 2010; Mayberry, Crocker, & Knoeferle, 2009; Knoeferle & Crocker, 2006, 2007). Among these cues are sentence-based ones such as case marking, verb meaning or temporal adverbs, but also extralinguistic cues such as contrast between objects, or information from action depictions. For instance, when a verb refers to an action, participants rapidly integrate the action (and its associated thematic role relations between two characters) and use it as a cue for anticipating upcoming role fillers when the sentence context is otherwise ambiguous regarding the referents' thematic role relations (Knoeferle, Crocker, Scheepers, & Pickering, 2005).

As already mentioned, actions are by no means the only cue that can rapidly affect real-time spoken sentence comprehension. A speaker's emotional expression can for example enhance a listener's visual attention to valence-matching event photographs as they are identified by a spoken sentence (Carminati & Knoeferle, 2013). However, seeing an action does appear to be of substantial importance in informing comprehension, and comprehenders may prioritize actions even when they are very infrequent (e.g., Abashidze, Knoeferle, & Carminati, 2013). In that context, Abashidze and colleagues have argued that recently-seen actions may be prioritized because they can be verified and have actually happened, while the absence of a visible action leads to uncertainty as to whether that action is going to be performed (even if such a "future" action occurs overall very frequently in the experimental context). On the other hand, our everyday conversations do include communication about absent actions and events; even concrete verbs do not invariantly reference actions and their associated patients in the immediate environment. Especially when compared to other contextual cues, the precise importance of visible and/or depicted actions for situated language comprehension thus remains an open issue.

Consider for instance, a speaker's eye gaze as another contextual cue for comprehenders. We know that when speakers talk about nearby objects, they tend to inspect them just 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 close link between speech-related eye movements and reference has important implications for language comprehension. Crucially, a speaker's gaze can help listeners to visually anticipate the next-mentioned referent (Hanna & Brennan, 2008; Knoeferle & Kreysa, 2012; MacDonald & Tatler, 2013; Staudte & Crocker, 2011).

In summary, numerous studies have shown that individual cues – such as actions or a speaker's eye gaze – can permit comprehenders to rapidly anticipate relevant referents. By contrast, only few studies have asked how the *type* of contextual cue affects visual anticipation. Neider, Chen, Dickinson, Brennan, and Zelinsky (2010), for instance, examined how two partners locate a randomly-appearing sniper target in a semi-realistic city environment. The two partners communicated either by shared voice, by shared gaze, or they could exploit both of these information sources in their joint search. Both partners had to locate the sniper target and make a joint decision. In the shared gaze condition, one partner would see the other's eye gaze in the form of a gaze cursor which was superimposed on the city scene. Partners took less time to find the target in the shared gaze than the shared voice condition, despite the arguably greater information content of the voice. This suggests that different cues can have differential benefits for successful communication between two interlocutors.

Another study pitted depicted but implausible action events against stereotypical thematic role knowledge associated with an agent. In the utterance *Den Piloten bespitzelt gleich…* ('The pilot (obj) spies-on soon'), the verb could either be grounded in a depicted spying action, thus guiding attention to its agent, or it could be related to a nearby stereotypical agent (a detective) depicted as performing an unrelated action. Faced with this ambiguity, participants preferred to inspect the agent of the depicted action over the stereotypical agent, prioritizing verb-action reference over expectations of what a stereotypical agent might do next (Knoeferle & Crocker, 2006).

Thus, with regard to allocating visual attention, people benefitted from relying on either gaze cues or depicted actions, relative to other contextual cues. But these cue preferences for speaker gaze on the one hand and depicted actions on the other hand emerged in two different tasks and experimental paradigms. A direct comparison of the two cues within the same paradigm and experiment is lacking.

One might argue that it is unsurprising that these two cues have distinct effects on visual attention. Real-world situations likely contain a myriad of cues to co-occurring speech content, among them information from the prior discourse, co-present objects, speaker gaze, and actions. These cues may appear sequentially, but often they will arguably all be available at the same time and compete for a comprehender's attention. Overt visual attention is by definition serial, and since it is a key player in relating language to the extralinguistic context, a better understanding of how comprehenders allocate visual attention to simultaneously competing cues is critical for models of visually-situated language comprehension.

This is because these models to date accommodate the rapid influence of *individual* contextual cues but – perhaps due to the lack of pertinent data – have neither modeled their combined effects nor do they say anything about the relative influence of different linguistic and non-linguistic cues. One notable exception is the model by Knoeferle and Crocker (2006; 2007), which proposes that reference to a depicted action by the verb has priority over associated world knowledge (which could prompt attention to a stereotypical role filler of an action). While they postulate a relative priority, other data suggest that distinct language-world relations are processed in a striklingly similar manner. Vissers et al. (2008), for instance, reported identical effects in event-related brain potentials in the processing of different kinds of spatial picture-sentence mismatches. Thus, the extent to which different contextual cues have distinct effects on a comprehender's visual attention remains a point of debate. Moreover, being able to quantify such informational biases is essential for refining current models. For instance, constraint-based models of language processing rely heavily on the probability of a cue for determining the strength of its influence on incremental language comprehension and ambiguity resolution (e.g., McRae, Tanenhaus, & Spivey, 1998). By contrast, they do not exploit other (informational) preferences that may guide (visual) attention and constrain comprehension. At least in part, this omission is due to the fact that little is known about how different non-linguistic cues compare in their effects on visual attention and language comprehension.

If several cues are available, they may frequently all point to the same referent as the most likely next-mentioned entity. Our study investigates this type of situation; we ask how two such cues conspire in enabling visual anticipation of referents, whether one disambiguating cue is more effective than another, and whether both cues jointly are more effective than a single cue in guiding visual attention. We also ask whether these cues differ in how they are inspected visually. To our knowledge, no such comparison has been reported to date, despite its relevance to modelling typical instances of situated language comprehension.

In order to address these open questions, we pitted two contextual cues against each other. Specifically, we compared speaker gaze with depicted actions in a design in which either one of these cues, both cues, or neither were available to comprehenders during utterance presentation. This allowed us to investigate whether the influence of the two cues on spoken comprehension is additive or interactive, as well as the extent to which they affect processing in a similar or different manner. In this context, it is important to note that the two cues differ in important ways: While speaker gaze shifts can be processed peripherally (at least to the extent that they are accompanied by head movements, as in our stimuli), benefitting from depicted actions generally requires both object recognition and semantic integration with the spoken sentence.

We recorded participants' fixations as they watched videos of a speaker producing a transitive sentence about two virtual characters. Post-sentence, participants verified whether a schematic depiction of role relations matched (vs. mismatched) the thematic role relations of the previously-heard sentence. Critically, we varied (a) whether the speaker shifted her gaze between the sentence referents, and (b) whether an object semantically related to the verb appeared between them. Differences in the effects of the two cues could reveal themselves in anticipatory fixation of the next-mentioned character and in post-sentence response times.

Eye-tracking Experiment

Methods

Participants Thirty-two Bielefeld University students participated in the experiment (ages 19-31). All were native speakers of German, had normal or corrected-to-normal vision and participated for course credit. All gave informed consent.

Materials and Design Using the virtual platform SecondLife®, we created 24 experimental and 48 similar filler items. For each item, we recorded a video of a speaker looking at three easily recognizable SecondLife® characters on a computer monitor. As the speaker inspected these characters, she produced a sentence describing an event taking place between two of them.

The experimental sentences were in German and all had a subject-verb-object structure (passive and dative-initial sentences occurred in some filler sentences; both are grammatical in German). Experimental trials depicted the agent in the centre of the scene and mentioned it first in the sentence. This is illustrated by the sentence *Der Kellner beglückwünscht den Millionär am Nachmittag* ('The waiter congratulates the millionaire in the afternoon') together with Figure 1 showing the waiter as the central on-screen character. The two characters situated to the left and right of the agent were the second-mentioned sentential patient (e.g., the millionaire on the right) and an unmentioned "competitor", respectively. Half the videos showed the patient on the right side of the screen, half on the left.

The items were assigned to eight lists in a 2×2×2 design: Speaker gaze was the first factor: In 50% of trials, the speaker was visible, in which case she shifted gaze from the agent to the patient character before mentioning the latter. In the other 50% of trials, a grey bar obscured the speaker.

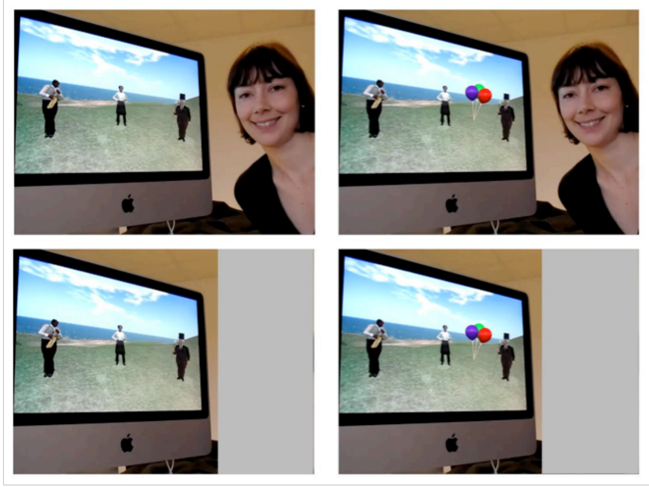

Figure 1: Examples of all four video conditions, clockwise from top left: Speaker-only, Speaker & Action, Action-only, no cue baseline.

The second factor was object-presence: In 50% of trials, an action-related object was presented on the screen exactly when the speaker began to shift gaze. This object related semantically to the action described by the verb (e.g., a bunch of balloons symbolized the verb "congratulate"). The object appeared between the agent character (e.g., the waiter) and the patient (the millionaire). Noticing the position of this action-related object was informative about the upcoming patient – similar to noticing the direction of the speaker's gaze shift. Figure 1 illustrates the four speaker × action conditions.

The third factor ('Match') related to the congruency of the sentence with a post-trial response template. In 50% of experimental trials an arrow pointed from the position of the waiter to that of the millionaire, thus matching the directionality of the sentential role relations; in the other 50% of cases it pointed from one of the two outer characters towards the central waiter, leading to a mismatch. The filler trials ensured that all four response templates occurred equally often, and that overall there was an equal number of matches and mismatches.

Procedure After participants had given informed consent, the eye-tracker (EyeLink 1000, SR Research) was set up for monocular tracking of the right eye with a 9-point calibration procedure. Participants received on-screen instructions and four practice trials. Each trial began with a drift correction, followed by the video. The video always showed the speaker smiling into the camera during the first few frames. Then she looked at the middle character, the right, and the left character, and back to the middle character (i.e. the agent) before beginning to speak. This inspection sequence was identical across all trials.

During the sentence, the speaker shifted gaze once more, turning her head to the right or left of the agent in order to look at the second-mentioned patient character. This gaze shift began just after the onset of the verb (*M* = 711 ms before the onset of the patient noun phrase). At the end of the sentence, the speaker looked back into the camera, the video terminated, and the response template appeared.

Participants' task was to watch the video, listen to the sentence, and then to indicate as quickly as possible by pressing one of two buttons on a Cedrus® response box whether the arrow on the response template correctly pointed from the position of the agent to the patient. For half of the participants the "match" response button was the left button on the button box; for the other half the button assignment was reversed. Participants took a short break after 36 trials, followed by a recalibration. At the end of the experiment, participants filled in a debrief questionnaire permitting us to assess whether they had guessed the purpose of the experiment.

Analyses We analyzed log-transformed response times (RTs) for accurate responses within 2 *SD* of each participant's mean per Match condition. For the analysis we used linear mixed models with crossed random intercepts and slopes for participants and items. Following Knoeferle and Kreysa (2012), fixation patterns were analyzed as mean log probability ratios for gazes to the patient character relative to the unmentioned competitor (*ln(P(patient)/ P(competitor)*)). A score of zero indicates equal attention to the patient and the competitor; a positive score implies the patient was fixated more, and a negative score that it was fixated less than the competitor. These log gaze probability ratios were computed for two time windows: The first

('SHIFT') spanned eight 100 ms time bins from the onset of the speaker's gaze shift (which was also the time at which the tool appeared), lasting roughly until the onset of the determiner of the patient noun phrase. The second time window ('NP2') comprised the first eight 100 ms bins from the onset of the patient noun phrase (about half its total duration). We fitted separate linear models for log ratios averaged over participants and items.

The initial models included three fixed factors for RTs (Match, Speaker, and Action) and three fixed factors for log-ratio gaze probabilities (Speaker, Action, and Time bin; Time bin had eight 100 ms-levels to capture developments across time), as well as all two-way interactions, random intercepts for participants and/ or items, and random slopes with the fixed factors and their interactions. This full model was fitted by maximum likelihood; in cases where it did not converge (this only ever occurred in RT analyses), interaction terms were removed from the random parts of the model in rising order of variance explained. The first converging model according to this strategy was defined as the maximal model, against which all simpler models were compared by log-likelihood ratio tests. We also ascertained via log-likelihood ratio tests whether interactions in the fixed-effects structure improved model fit for the maximal compared to simpler models. Fixed-effect interactions that did not contribute significantly were removed, as were the corresponding random slopes, until model fit either did not improve further, or until a main-effects-only model remained. In this final model, we again included as many random slopes corresponding to the fixed effects as possible, while maintaining convergence. We report the *t*-values for all fixed effects and interactions in the final models. Following standard procedure in the literature, we considered coefficients as significant only if the absolute value of the *t*-statistic exceeded 2. We report the *t*-values.

Results

Response times We calculated the time from the onset of the response template until the button press for correct responses and analyzed the log-transformed RTs. The fixed part of the final model for RTs contained only the three main effects of Match, Speaker, and Action; the random part consisted of the random intercepts for participants and items, and a random slope each for Match, Speaker, and Action by participants only (R^2 = .505, sigma = 0.251). In this model, only the factor Match affected RTs (t = -5.40). Participants verified a match between sentence and template faster by around 150 ms (M = 927 ms) than a mismatch (M = 1078 ms).

Eye-movement analyses We compared the allocation of attention to the patient character over time between the conditions (combined cues; speaker-only; action-only; no-cue). Figure 2 graphs the fixation patterns to the patient character, beginning when the speaker shifted her gaze or the object appeared. Since the videos did not differ before that point in time, this was the first opportunity at which participants' visual attention to the patient could be affected by the two types of cues.

Figure 2 shows an early increase of fixations to the patient character in the speaker-only condition (solid red line): By about 500 ms after gaze shift, participants had followed the speaker's gaze to the patient, which was well before this character was mentioned.

By contrast, for the conditions in which an action-related tool was displayed, listeners' attention shifted to the patient only later, while it was being named (the purple and blue lines). This delay is likely due to an abrupt drop in patient fixations just after the action tool appeared. At this point, participants' attention was drawn to the onsetting tool. Finally, when no cue was available (dotted black line), fixations to the patient increased only once the patient had been mentioned (as expected for the baseline).

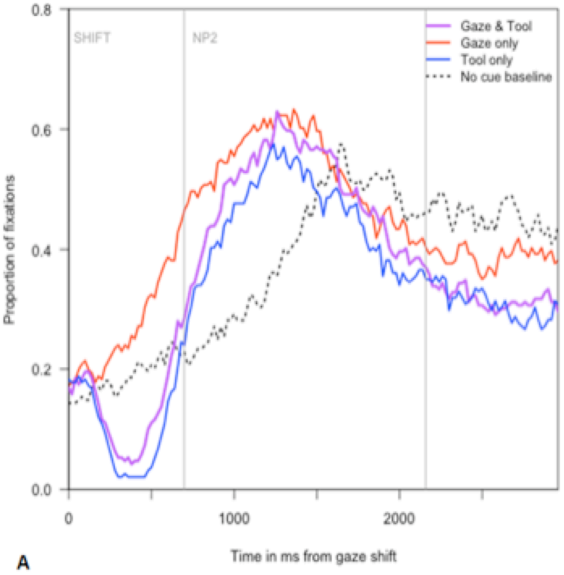

Figure 2: Proportion of fixations over time to the patient character, by condition. The graph begins at the onset of the speaker's gaze shift, which was also the onset of the action object, if visible. Mean onsets of the NP2 and of the ending phrase are marked by grey vertical bars.

Inferential analyses of log gaze probability ratios corroborated the descriptive impression: In the earlier SHIFT time window, participants generally fixated the patient more than the competitor (*t*s > 3.7), and this tendency increased over time (*t*s > 8). Both Speaker gaze and the Action tool increased the likelihood of fixating the patient in the by-items analysis, (*t*s > 2.4), but this was not apparent in the by-participants analysis, nor did the two factors interact. However, both Speaker (by participants) and Action interacted with the Time bin factor, reflecting a general increase in patient fixations over time.

In the NP2 time window, once the speaker began to speak about the patient, participants were even more likely to fixate this character than the competitor (*t*s > 17). More interestingly, this tendency increased substantially both when they had just seen the speaker's gaze shift (*t*s > 6), and when an action-related object had appeared on the screen (*t*s > 4). In this time window, Speaker interacted with Action such that the combined availability of both cues led to *less* patient fixations than the gaze-only condition, but to more patient fixations than action-only (*t*s = |5|)). Only when neither cue was present were participants equally likely to fixate the patient and the competitor character.

General Discussion

We pitted two types of contextual information against each other, both of which have individually been shown to facilitate spoken sentence comprehension: In one condition, participants could see how a speaker shifted gaze to look at a depicted character she was about to mention – thus, gaze provided a cue for listeners to predict how she would continue her sentence. In the contrasting condition, an object related to the verb appeared on-screen. In this case, listeners were able to predict the next-to-be-mentioned patient by integrating the identity of this object with the meaning of the verb. We also included a condition in which both types of cues were available simultaneously, as well as a baseline condition with no predictive cues.

In all three conditions with predictive cues, listeners were significantly faster to fixate the upcoming patient referent of the sentence than in the no-cue baseline condition. They were able to use either type of cue to guide visual attention and arguably to anticipate that this character would be mentioned next.

At the same time, we found interesting disparities in how each of these two contextual cues affected immediate fixation patterns during sentence processing. Seeing the speaker's gaze shift in the absence of an action object led listeners to follow this gaze shift to the patient almost immediately and without directly fixating the speaker. This suggests the implication of low-level, potentially peripheral processing of the gaze and head shift. In contrast, although appearing action objects ultimately enabled participants to fixate the upcoming referent roughly to the same extent as did the gaze shift, the action-based anticipatory fixations occurred approximately 200 ms later, due to prior fixation of the action object itself. Interestingly, the simultaneous availability of both types of cues was at most as helpful as the gaze cue on its own, never better.

We can draw two important conclusions from these results: First, not all contextual cues are equal with regard to how they influence ongoing language comprehension. Here, further research is necessary to explore the nature of these (and other) distinct cues, with the aim of extending existing models of comprehension with an account of both the relative and joint effects of distinct visual cues. Second, more is not always better: It seems that contextual cues have a ceiling in the facilitation they can provide. A single cue can be sufficient to achieve this level of facilitation; additional cues do not necessarily lead to further benefit.

One potential limitation of the design used here is that it was possible for listeners to use the appearing action object without considering its identity, since it always appeared between the agent character and the upcoming patient. Thus, independent of the semantic content of the depicted object, its location on the screen pointed unambiguously to the next-to-be-mentioned character. In consequence, fixating the patient character did not necessarily require participants to process the identity of the object and to match it semantically to the verb they had just heard. We aim to clarify this issue in a future study by having an action tool appear on either side of the agent (see Kreysa, Nunnemann, & Knoeferle, 2013, for preliminary results). We will present a verb-irrelevant object between the agent and the competitor character, and a verb-related object between agent and patient character, making it essential to check the semantics of the depicted object(s) with the semantics of the verb. It is possible that such verb-action integration takes time and that performing such integration would further delay the visual anticipation of the target character relative to the gaze condition. Alternatively, it is possible that even in the present study, where only one object appeared, participants did integrate the verb with the action object. In this case, we should see no substantial difference in visual anticipation even when two competing action objects appear (one on either side of the agent).

In spite of this potential limitation, the present study shows clearly that a single contextual cue is all it takes to predict upcoming sentence content; more cues do not result in greater facilitation. In addition, the easier or more superficial a particular cue is to process, the faster its effect is on fixation patterns. An open question, however, is whether faster anticipation necessarily means better understanding. While prediction may benefit comprehension, in-depth processing and encoding of information about the patient arguably requires more than a few rapid fixations to that character. In future research, we plan to include a memory gating task to address this issue. Post-experiment, participants will be asked to sequentially recall components of the sentence heard during the main experiment (first the verb, then the patient of the sentence, assisted by images of the action tool and potential patient characters, respectively). If either or both of our contextual cues affect participants' short-term (post-experiment) memory of different sentence components, then this should be reflected in the recall rates. For instance, if the action (but not gaze) is truly integrated with the verb, then we should see better recall of sentence content if participants saw an action-related object in the main experiment than if they saw only the speaker or neither cue. By contrast, if gaze is particularly useful in cueing visual attention to, and subsequent encoding of upcoming referents in memory, then recognition of the patient in the gating task should be better for the gaze than for the action conditions.

Meanwhile, the present results already reveal that contextual aspects of situated language can differ in the time course with which they affect language-mediated attention. We believe that this important fact has so far received too little attention in theories and models of situated sentence comprehension.

Acknowledgments

This research was funded by the Cognitive Interaction Technology Excellence Center (German Research Foundation, DFG) and by a grant within the SFB 673 “Alignment in Communication” awarded to PK (DFG).

References

Abashidze, D., Knoeferle, P., & Carminati, M. N. (2013). Gaze cue effect during language comprehension. In: R. Fernández & A. Isard (Eds). *Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue*. Amsterdam.

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

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

Bock, K., Irwin, D. E., Davidson, D. J., & Levelt, W. J. M. (2003). Minding the clock. *JML, 48*, 653-685.

Carminati, M. N., & Knoeferle, P. (2013). Effects of speaker emotional facial expression and listener age on incremental sentence processing. *PLoS ONE 8*(9): e72559. doi:10.1371/ journal.pone.0072559.

Crocker, M. W., Knoeferle, P., & Mayberry, M. (2010). Situated sentence processing: The coordinated interplay account and a neurobehavioural model. *Brain and Language, 112*, 189-201.

Griffin, Z. M. (2004). Why look? Reasons for eye movements related to language production. In J. M. Henderson & F. Ferreira (Eds.), *The Interface of Language, Vision, and Action: Eye Movements and the Visual World* (pp. 213-247). New York: Psychology Press.

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

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

Huettig, F., Rommers, J., & Meyer A. (2011). Using the visual world paradigm to study language processing. *Acta Psychologica, 137*, 151-171.

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

Knoeferle, P., & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: evidence from eye movements. *JML, 57*, 519-543.

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

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

Kreysa, H., Nunnemann, E. M., & Knoeferle, P. (2013). Comparing effects of speaker gaze and action information on anticipatory eye movements during spoken sentence comprehension. In K. Holmqvist, F. Mulvey & R. Johansson (Eds.), Book of Abstracts of the 17th European Conference on Eye Movements, Lund, Sweden. *Journal of Eye Movement Research, 6*(3), 150.

Kuchinsky, S. E., Bock, K., & Irwin, D. E. (2011). Reversing the hands of time: Changing the mapping from seeing to saying. *JEP: LMC, 37*, 748-756.

MacDonald, R. G., & Tatler, B. W. (2013). Do as eye say: Gaze cueing and language in a real-world social interaction. *Journal of Vision, 13*(4):6, 1-12.

Mayberry, M., Crocker, M. W., & Knoeferle, P. (2009). Learning to Attend: A Connectionist Model of the Coordinated Interplay of Utterance, Visual Context, and World Knowledge. *Cognitive Science 33,* 449-496.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38*, 283–312.

Meyer, A. S., & Lethaus, F. (2004). The use of eye tracking in studies of sentence generation. In J. M. Henderson & F. Ferreira (Eds.), *The Interface of Language, Vision, and Action: Eye Movements and the Visual World* (pp. 191-211). New York, Hove: Psychology Press.

Meyer, A. S., Roelofs, A., & Levelt, W. J. M. (2003). Word length effects in object naming: The role of a response criterion. *Journal of Memory and Language, 48*, 131-147.

Neider, M. B., Chen, X., Dickinson, C. A., Brennan, S. A., and Zelinsky G. J. (2010). Coordinating spatial referencing using shared gaze. *Psychonomic Bulletin & Review 17* (5), 718-724. doi:10.3758/PBR.17.5.718

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

Vissers, C., Kolk, H., Van de Meerendonk, N., & Chwilla, D. (2008). Monitoring in language perception: Evidence from ERPs in a picture-sentence matching task. *Neuropsychologia, 46,* 967–982.