# Using listener gaze to augment speech generation in a virtual 3D environment

Maria Staudte
Saarland University

Alexander Koller
University of Potsdam

Konstantina Garoufi
University of Potsdam

Matthew Crocker
Saarland University

**Abstract**

Listeners tend to gaze at objects to which they resolve referring expressions. We show that this remains true even when these objects are presented in a virtual 3D environment in which listeners can move freely. We further show that an automated speech generation system that uses eyetracking information to monitor listener's understanding of referring expressions outperforms comparable systems that do not draw on listener gaze.

## Introduction

In situated spoken interaction, there is evidence that the gaze of interlocutors can augment both language comprehension and production processes. For example, speaker gaze to objects that are about to be mentioned (Griffin & Bock, 2000) has been shown to benefit listener comprehension by directing listener gaze to the intended visual referents (Hanna & Brennan, 2007; Staudte & Crocker, 2011; Kreysa & Knoeferle, 2011). Even when speaker gaze is not visible to the listener, however, listeners are known to rapidly attend to mentioned objects (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). This gaze behavior on the part of listeners potentially provides speakers with useful feedback regarding the communicative success of their utterances: By monitoring listener gaze to objects in the environment, the speaker can determine whether or not a referring expression (RE) they have just produced was correctly understood or not, and potentially use this information to adjust subsequent production.

In this paper we investigate the hypothesis that speaker use of listener gaze can potentially enhance interaction, even when situated in complex and dynamic scenes that simulate physical environments. In order to examine this hypothesis in a controlled and consistent manner, we monitor listener performance in the context of a computer system that generates spoken instructions to direct the listener through a 3D virtual environment with the goal of finding a trophy. Successful completion of the task requires listeners to press specific buttons. Our experiment manipulated whether or not the computer system could follow up its original RE with feedback based on the listener's gaze or movement behavior, with the aim of shedding light on the following two questions:

- Do listener eye movements provide a consistent and useful indication of referential understanding, on a per-utterance basis, and when embedded in a dynamic and complex, goal-driven scenario?
- What effect does gaze-based feedback have on listeners' (gaze-)behavior and does it increase the more general effectiveness of an interaction?

We show that the listeners' eye movements are a reliable predictor of referential understanding in our virtual environments. A natural language generation (NLG) system, that exploited this information to provide direct feedback, communicated its intended referent to the listener more effectively than similar systems that did not draw on listener gaze. Gaze-based feedback was further shown to increase listener attention to potential target objects in a scene, indicating a generally more focused and task-oriented listener behavior. This system is, to our knowledge, the first NLG system that adjusts its referring expressions to listener gaze.

### Related work

Previous research has shown that listeners align with speakers by visually attending to mentioned objects (Tanenhaus et al., 1995) and, if possible, to what the speaker attends to (Richardson & Dale, 2005; Hanna & Brennan, 2007; Staudte & Crocker, 2011). Little is known, however, about speaker adaptation to the listener's (gaze) behavior, in particular when this occurs in dynamic and goal-oriented situations. Typically, Visual World experiments have used simple and static visual scenes and disembodied utterances and have analyzed the recorded listener gaze off-line (e.g., Altmann & Kamide, 1999; Knoeferle, Crocker, Pickering, & Scheepers, 2005). Although studies involving an embodied speaker inherently include some dynamics in their stimuli, this is normally constrained to speaker head and eye movements (Hanna & Brennan, 2007; Staudte & Crocker, 2011). Besides simplifying the physical environment to a static visual scene, none of these approaches can capture the reciprocal nature of interaction. That is, they do not take into account that the listeners' eye movements may, as a signal of referential understanding to the speaker, change the speaker's behavior and utterances on-line and, as such, affect the listener again.

One study that emphasized interactive communication in a dynamic environment was conducted by Clark and Krych (2004). In this experiment, two partners assembled Lego models: The directing participant advised the building participant on how to achieve that goal. It was manipulated whether or not the director could see the builder's workspace and, thus, use the builder's visual attention as feedback for directions. Clark and Krych found, for instance, that the visibility of the listener's workspace led to significantly more deictic expressions by the speaker and to shorter task completion times. However, the experimental setting introduced large variability in the dependent and independent variables, making controlled manipulation and fine-grained observations difficult. In fact, we are not aware of any previous work that has successfully integrated features of natural environments—realistic, complex and dynamic scenes in which the visual salience of objects can change as a result of the listener's moves in the environment—with the reciprocal

nature of listener-speaker-adaptation while also being able to carefully control and measure relevant behavioral data.

Recently, researchers have examined eye gaze of speakers and listeners in the scenes of Tangram puzzle simulations on computer screens (Kuriyama et al., 2011; Iida, Yasuhara, & Tokunaga, 2011). In these experiments, eye gaze features are found to be useful for a machine learning model of reference resolution. However, this setting is restricted in its dynamics, as it does not embed the objects into physical scenes or involve any updates to the spatial and visual context of the objects in the scenes. In contrast, by *generating* REs and asking the subjects to resolve them, rather than resolving human-produced REs itself, the system we propose here can provide more control over the language that is used in the interaction.

Computational models of gaze behavior are frequently implemented in embodied conversational agents as part of non-verbal behavior that aims at improving the human-computer interaction (see e.g. Foster, 2007). Such agents do not typically employ listener gaze tracking for the generation of appropriate REs, though. One work that focuses on situated RE generation is Denis (2010), which takes the visual focus of objects into account for the gradual discrimination of referents from distractors in a series of utterances. However, visual focus in Denis' work is modeled by visibility of objects on screen rather than eye gaze. To our knowledge, there exists no prior RE generation algorithm that is informed directly by listener gaze.

Finally, gaze as a modality of interaction has been investigated in virtual reality games before, e.g. by Hülsmann, Dankert, and Pfeiffer (2011). However, most such settings do not use language as a further modality. One virtual game-like setting which focuses on language is the recent Challenge on Generating Instructions in Virtual Environments (GIVE; Koller et al., 2010), which evaluates NLG systems that produce natural-language instructions in virtual environments. In this work we use the freely available open-source software infrastructure provided by GIVE[1] to set up our experiment.

## Methods

In the GIVE setting (Koller et al., 2010; Striegnitz et al., 2011), a human user can move about freely in a virtual indoor environment featuring several interconnected corridors and rooms. A 3D view of the environment is displayed on a computer screen as in Fig. 1, and the user can walk forward and backward, and turn left and right, using the cursor keys. They can also navigate to buttons and, once they have approached them closely enough, click on them with the mouse to press them. In Fig. 1 the object currently under inspection by the user is the rightmost button on the wall, marked with a large white circle. The trace of the fixation's coordinates is rendered by smaller white circles. These gaze markings do not appear on the user's screen during the experiment.

The user interacts with an NLG system in the context of a treasure-hunt game, where the user's task is to find a trophy hidden in a wall safe. They must press certain buttons in the correct sequence in order to open the safe; since they do not have prior knowledge of which buttons to press, they rely on instructions and REs generated by the NLG system in order to carry out the task. A room may contain several buttons other than the *target*, which is the button that the user must press next. These other buttons are called *distractors* and are there to make the RE resolution task more challenging. Rooms also contain a number of landmark objects, such as chairs and plants, which cannot be interacted with, but may be used in REs to nearby targets. For our experiment we use three different virtual environments designed by Gargett, Garoufi, Koller, and Striegnitz (2010), which differ in what objects they contain and where they are located.


Figure 1: A screenshot of one of the virtual 3D environments.

### Generation systems

We implemented three different NLG systems for generating instructions in these virtual environments. All systems generate navigation instructions, which guide the user to a specific location, as well as object manipulation instructions such as "press the blue button" containing REs such as "the blue button". The generated instructions are converted to speech by the MARY text-to-speech system (Schröder & Trouvain, 2003) and presented via loudspeaker. At any point, the user may press the 'H' key on their keyboard to indicate that they are confused. This will cause the NLG system to generate a clarified instruction. All three systems operate on the same codebase for the generation of simple yet effective navigation instructions (e.g. "go through the doorway"), but differ in their RE generation strategies.

Our baseline system generates REs that are optimized for being easy for the listener to understand, according to a corpus-based model of understandability (Garoufi & Koller, 2011). Crucially, this system does not monitor whether the listener understood an RE. It never gives any (positive or negative) feedback, and will only generate a follow-up RE if the user either asks for help ('H' key) or presses the wrong button. Therefore we call this system the *no-feedback* system.

The *movement* system extends the no-feedback system by

[1] http://www.give-challenge.org/research

monitoring the user's movements in the game after it has uttered an RE, and attempting to predict whether they will press the button it described or not. This system does nothing until only a single button in the current room is visible to the user; then it tracks the user's distance from this button, where "distance" is a weighted sum of walking distance to the button and the angle the user must turn to face the button. If, after hearing the RE, the user has decreased the distance by more than a given threshold, the system concludes that the hearer has resolved the RE as this button. If it is the button the system intended to refer to, it utters the positive feedback "yes, that one!" For incorrect buttons, it utters the negative feedback "no, not that one."

Finally, the *eyetracking* generation system attempts to predict whether the user will press the correct button or not by monitoring their gaze. At intervals of approximately 15 ms, the system samples the (x,y) position on the screen that the user is looking at. It then resolves this (x,y) screen position to an object in the 3D scene. If the user fixates the same object for more than 300 ms, the system counts this as an inspection of that object; interruptions of the inspection of less than 150 ms are ignored. Once it has detected an inspection to a button in the room, the eyetracking system generates positive or negative feedback utterances in exactly the same way as the movement system does.

The system maps the screen positions reported by the eyetracker to 3D objects as follows: When the 3D engine renders the 3D scene onto the 2D screen, it assumes a certain position of the "camera" in the 3D environment; this roughly corresponds to the position of the user's eyes. For each object that is currently visible, the system computes its bounding box, i.e. the smallest box that completely contains the object. It determines the minimum angle $\alpha$ between the ray from the camera position to some corner of the bounding box and the ray from the camera position to the center of the bounding box. Intuitively, $\alpha$ represents the size of the object on the screen. The system also determines the angle $\beta$ between the ray from the camera position to the (x,y) position in the screen plane reported by the eyetracker and the center of the bounding box. Small values of $\beta$ represent situations in which the user looks directly at the center of an object. An object is a candidate for being fixated if one of $\beta/\alpha$ or $\beta - \alpha$ is below a certain threshold. Among all candidates (if there are any), the system then finally chooses the object with the smallest $\beta$.

Both the movement-based and the eyetracking-based model withhold their feedback until a first full description of the referent (a *first-mention RE*) was spoken. Additionally, they only provide feedback on newly approached or inspected buttons and will not repeat this feedback unless the listener has approached or inspected another button in the meantime. We call the time between the onset of the first-mention RE and the next button press in a scene, the *critical time region*.

## Participants

Thirty-one students, enrolled at Saarland University, were paid to take part in this study (12 females). All reported their English skills as fluent, and all were able to complete the tasks. Their mean age was 27.6.

## Task and procedure

A faceLAB eyetracking system[2] remotely monitored participants' eye movements on a 24-inch monitor. Before the experiment, participants received written instructions that described the task and explained that they would be given instructions by an NLG system. They were encouraged to request additional help anytime they felt that the instructions were not sufficient (by pressing the 'H' key).

The eye-tracker was calibrated using a nine-point fixation stimulus. We disguised the importance of gaze from the participants by telling them that we videotaped them and that the camera needed calibration. Participants then started with a short practice session to familiarize themselves with the game controls and to clarify remaining questions, before playing three full games (each with a different virtual environment and generation system). The order of games was alternated according to the Latin square design. Finally, each participant received a questionnaire which aimed to assess whether participants noticed that they were eye-tracked and that one of the generation systems made use of that. The entire experiment lasted approximately 30 minutes.

## Analysis

Firstly, we determined whether the participant pressed the correct button (without having to ask for help by pressing the 'H'-key) by comparing each button the participant pressed with the target referent of the most recent first-mention RE. REs that did not lead to a button press (e.g. because the participant navigated away to another room, causing the system to switch to navigation instructions) were considered unsuccessful. This served as a dependent variable but also as a means for subdividing data according to un-/successful trial completion. Secondly, inspections recorded on a button in the player's room, i.e., on the target or a distractor, during the critical time region were registered in all conditions (not just the eyetracking NLG system) and analyzed as a main dependent variable. Further total trial time, i.e., the time taken from the onset of an RE to the button press, as well as the onset time of system feedback (when provided) were recorded. Finally, we considered the frequency with which participants asked for help by pressing the 'H' key as a measure of confusion.

To control for external factors, we discarded individual scenes in which the systems rephrased their first-mention REs (e.g. by adding further attributes), as well as a few scenes which the participants had to go through a second time due to technical glitches. To remove errors in eyetracker calibration, we included interactions with the eyetracking NLG system in the analysis only when we were able to record inspections (to the referent or any distractor) in at least 80% of all referential scenes. This filtered out 9 interactions out of the 93 we collected.

[2] http://www.seeingmachines.com/product/facelab

Inferential statistics on this data were carried out using mixed-effect models from the lme4 package in R (Baayen, Davidson, & Bates, 2008). Specifically, we used logistic regression for modeling binary data such as referential success rates, Poisson regression for count variables (e.g., 'H'-key strokes) and linear regression for inspection durations. Further, main effects and interactions were determined through model reduction, which assesses the contribution of a predictor or interaction to a fitting model by running a $\chi^2$-comparison between models with and without the particular predictor(s).

## Results

The post-task questionnaires, revealed no differences in participants' preferences for any particular NLG system. Similar numbers of participants chose each of the systems on questions such as "which system did you prefer". When asked for differences between the systems in free-form questions, no subject mentioned eye gaze. We take this to mean that the participants did not realize they were being eyetracked.

### Eye movements

We recorded and analyzed inspections to target and distractor buttons in all conditions. Mean inspection durations during the critical time region (reference onset until button press) were correlated with the success in pressing the correct button and are provided in Table 1.

To investigate our first hypothesis, namely that listener eye movements provide a consistent and useful indication of referential understanding even when embedded in a dynamic, complex and goal-driven scenario, we first consider our baseline condition, the no-feedback system, separately: Model reduction revealed that both inspection duration on the target and inspection duration on the distractors indeed predict success ($\chi^2(1) = 28.87, p < .001$ and $\chi^2(1) = 96.24, p < .001$, respectively). While target inspection duration positively predicts success (Coeff. = 0.00110, SE = 0.00024, Wald's Z = 4.53 , $p < .001$), distractor inspections negatively predict success (Coeff. = $-0.00178$, SE = 0.00027, Wald's Z = $-6.71, p < .001$).

Further, to assess the influence of gaze-based feedback back on listeners' gaze behavior, we investigated whether the type of system used for generating REs did in fact influence inspection durations (as given in Table 1). We fitted models to target inspection duration and distractor inspection duration using system as predictor, for successful and unsuccessful scenes separately. Model reductions revealed a main effect of system (target: $\chi^2(2) = 12.79, p < .01$, distractor: $\chi^2(2) = 47.10, p < .001$) on both inspection variables, but only in successful scenes. That is, with the eyetracking-based feedback system, participants inspected both the target and distractor buttons longer than with the other two systems. An average trial also lasted longer with this system than with the no-feedback system. In unsuccessful scenes no significant differences between inspection durations were observed.

Table 1: Mean inspection durations for target and distractor buttons and the total trial time in milliseconds, for successful and unsuccessful button presses separately. (ET = eyetracking-based system, MOV = movement-based system, NO = no-feedback system.) Differences to ET are significant at: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, # $p < 0.1$.

| System (# Trials) | Target | Distractor | Trial Total |
|---|---|---|---|
| Successful: | | | |
| **ET** (182) | 2111.6 | 720.5 | 8096 |
| **MOV** (258) | 1493.8** | 260.5*** | 7418 |
| **NO** (237) | 1492.0*** | 185.7*** | 6877** |
| Unsuccessful: | | | |
| **ET** (16) | 752.1 | 3378.9 | 10892 |
| **MOV** (37) | 602.6 | 2113.1 | 10343 |
| **NO** (47) | 619.5 | 1891.7 | 9130 |

However, this is most likely due to the low amount of unsuccessful scenes.

Finally, we considered only cases in which feedback was indeed given in order to more precisely assess the influence of effective feedback (types) on participant inspections during reference resolution. Table 2 shows this data further subdivided into scenes with initially positive feedback and scenes with initially negative feedback (the eyetracking system is used as intercept for comparisons between both systems). This is to explore the effect of positive and confirming feedback given by each system and the possibly different effect of negative feedback which unspecifically re-directs the participant to other buttons. We observed that positive feedback of both systems leads to a similar increase of target and decrease of distractor inspections (cf. before and after columns in Table 2). However, eyetracking-based feedback was given earlier (Coeff. = 573.6, SE = 240.2, t = 2.39, $p(\text{MCMC}) < 0.05$) and led to overall longer inspections of the target *and* distractor buttons relative to the trial duration. That is, participants spent significantly more time of a trial (34.1%) looking at potential target buttons than with movement-based feedback (25.5%, Coeff. = $-0.0552$, SE = 0.0178, t = $-3.11, p(mcmc) < 0.01$). This effect was even larger with negative feedback where the difference in feedback onset was even greater (Coeff. = 1237.8, SE = 378.1, t = 3.27, $p(\text{MCMC}) < 0.01$) and the relative button inspection time was also longer (Coeff. = $-0.1818$, SE = 0.0283, t = $-6.43, p(\text{MCMC}) < 0.001$). Possibly because of this large difference in feedback onset, we also found (marginally) longer inspections to the buttons after feedback onset.

### Interaction Effectiveness

To evaluate our second hypothesis, namely that gaze-based feedback potentially sustains a more effective interaction than other or no feedback, we considered several indicators for in-

Table 2: Mean values for initial positive and negative feedback separately: inspection durations for target and distractor buttons (before and after feedback onset), feedback onset times, total trial durations, proportion of time spent fixating buttons during trials, and referential success rates. Differences to ET are significant at: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, # $p < 0.1$.

| | Target | | Distractor | | Feedback Onset | Trial Total | Button Fix. Proportion | Success |
|---|---|---|---|---|---|---|---|---|
| | Before | After | Before | After | | | | |
| Positive Feedback: | | | | | | | | |
| ET | 513 | 1389 | 111 | 67 | 4115 | 6511 | 34.1 | 97.6 |
| MOV | 465 | 1123 | 196 | 30 | 4688* | 7051* | 25.5** | 97.0 |
| Negative Feedback: | | | | | | | | |
| ET | 109 | 2155 | 733 | 1596 | 3987 | 11888 | 39.5 | 84.0 |
| MOV | 120 | 926*** | 484# | 802# | 5225** | 11319 | 20.1*** | 68.0* |

teraction effectiveness. As a first measure, we looked at the frequency with which participants pressed the 'H' key to indicate their confusion. The overall average of 'H' keystrokes per game was 1.14 for the eyetracking generation system, 1.77 for the movement system was employed, and 2.26 for the no-feedback system. A model fitted to the key stroke distribution per system shows significant differences both between the eyetracking and the no-feedback system (Coeff. = 0.703, SE = 0.233, Wald's Z = 3.012, $p < .01$) and between the eyetracking and the movement-based system (Coeff. = 0.475, SE = 0.241, Wald's Z = 1.967, $p = .05$).

A second measure of interaction quality is the ratio of all REs that the participants resolved correctly. Mean success rates for trials with feedback only are further reported in the final column of Table 2. Logistic mixed-effects models revealed a significant difference in success rates (Coeff. = $-0.918$, SE = 0.461, Wald's Z = $-1.990$, $p < .05$) for negative feedback while the success rates were similar for positive feedback. Additionally, total trial time is significantly shortened by positive (but not negative) eyetracking-based feedback (Coeff. = 713.7, SE = 311.4, t = 2.29, $p < .05$). Thus, when positive feedback was given, the eyetracking system had shorter trial times (along with earlier feedback), while having similar success rates as the movement system. Conversely, negative feedback led to similar trial times but with higher success rates by the eyetracking system.

## Discussion

Concerning our first hypothesis—that gaze reflects online referential understanding even in dynamic 3D environments—we find that participants indeed tend to rapidly fixate the object described by the system. Appropriate feedback by the eyetracking system, in turn, elicits longer inspection durations on potential targets, showing more focused, task-oriented listener attention.

This positive finding is further supported by the performance of the eyetracking NLG system, which outperforms the no-feedback baseline on listener confusion and on RE success rate. If gaze was not a reliable indicator of RE interpretation, this system would frequently give misleading feedback and therefore perform worse. Together with the finding that positive gaze-based feedback leads to shorter trial times than positive movement-based feedback, while negative gaze-based feedback leads to better success rates than negative movement-based feedback, this confirms our second hypothesis. That is, the eyetracking system (positively) influences interaction effectiveness.

One observation from the games in the experiment is that listeners tend to rapidly look back and forth between different buttons when they are confused. However, it needs to be still worked out, how to interpret such signals more generally. A further issue is that all objects in the 3D world shift on the screen when the user turns or moves in the virtual environment. The user's eyes will typically follow the object they are currently inspecting, but lag behind until the screen comes to a stop again. One topic for future work would be to remove such noise from the eyetracking signal.

Finally, the negative feedback our systems gave was very unspecific ("no, not that one", even when there were other distractors) and given earlier and numerically also more frequently by the eyetracking system. This could explain the different effects of positive and negative feedback on inspection behavior and the time-accuracy trade-off for each system: Longer trial times but better success rates for negative gaze-based (compared to movement-based) feedback. We used negative feedback to keep the experimental situation more controlled but the performance of the feedback systems could possibly be improved by giving more specific feedback ("no, the BLUE button"). Another avenue for future research is to examine whether listener gaze could also be useful for other NLG or dialog tasks apart from RE generation.

## Conclusion

We reported on an experiment in which an NLG system used listener gaze to track the listener's understanding of REs and provide positive or negative feedback when needed. This shows that listener gaze provides consistent and useful feedback about the listener's interpretation process, and that NLG systems can be improved by tracking this interpretation process in real time.

These findings have consequences both for psycholinguistics and for computational linguistics. On the psycholinguistic side, they open the way for eyetracking experiments that are set in a more natural and dynamic, and importantly, truly interactive, environment than traditional Visual World experiments. On the computational side, they offer a testbed for interactive NLG and dialogue systems; even though eyetracking devices are not yet commonplace as computer peripherals they can still allow us to implement and test theories of how to effectively track the comprehension process of the user.

## Acknowledgments

The research reported of in this paper was partly supported by the "Multimodal Computing and Interaction" Cluster of Excellence at Saarland University. We thank Irena Dotcheva for help with data collection as well as Alexandre Denis and Christoph Clodo for software support.

## References

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

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

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*(1), 62–81.

Denis, A. (2010). Generating referring expressions with reference domain theory. In *Proceedings of the 6th International Natural Language Generation Conference.*

Foster, M. E. (2007). Enhancing human-computer interaction with embodied conversational agents. In *Proceedings of HCI International 2007.*

Gargett, A., Garoufi, K., Koller, A., & Striegnitz, K. (2010). The GIVE-2 Corpus of Giving Instructions in Virtual Environments. In *Proceedings of the 7th Conference on International Language Resources and Evaluation.*

Garoufi, K., & Koller, A. (2011). The Potsdam NLG systems at the GIVE-2.5 Challenge. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation (ENLG).*

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

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

Hülsmann, F., Dankert, T., & Pfeiffer, T. (2011). Comparing gaze-based and manual interaction in a fast-paced gaming task in virtual reality. In *Virtuelle & Erweiterte Realität, 8. Workshop der GI-Fachgruppe VR/AR.*

Iida, R., Yasuhara, M., & Tokunaga, T. (2011). Multi-modal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In *Proceedings of 5th International Joint Conference on Natural Language Processing.*

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

Koller, A., Striegnitz, K., Byron, D., Cassell, J., Dale, R., Moore, J., et al. (2010). The First Challenge on Generating Instructions in Virtual Environments. In E. Krahmer & M. Theune (Eds.), *Empirical Methods in Natural Language Generation* (pp. 337–361). Springer.

Kreysa, H., & Knoeferle, P. (2011). Peripheral speaker gaze facilitates spoken language comprehension: syntactic structuring and thematic role assignment in German. In B. Kokinov, A. Karmiloff-Smith, & N. Nerssessian (Eds.), *Proceedings of the European Conference on Cognitive Science 2011.*

Kuriyama, N., Terai, A., Yasuhara, M., Tokunaga, T., Yamagishi, K., & Kusumi, T. (2011). Gaze matching of referring expressions in collaborative problem solving. In *Proceedings of International Workshop on Dual Eye Tracking in CSCW (DUET).*

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, *29*(6), 1045–1060.

Schröder, 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*, 365–377.

Staudte, M., & Crocker, M. W. (2011). Investigating joint attention mechanisms through human-robot interaction. *Cognition*, *120*(2), 268–291.

Striegnitz, K., Denis, A., Gargett, A., Garoufi, K., Koller, A., & Theune, M. (2011). Report on the Second Second Challenge on Generating Instructions in Virtual Environments (GIVE-2.5). In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation.*

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634.