Enhancing Referential Success by Tracking Hearer Gaze

Alexander Koller
University of Potsdam
koller@ling.uni-potsdam.de

Konstantina Garoufi
University of Potsdam
garoufi@uni-potsdam.de

Maria Staudte
Saarland University
masta@coli.uni-saarland.de

Matthew Crocker
Saarland University
crocker@coli.uni-saarland.de

Abstract

The ability to monitor the communicative success of its utterances and, if necessary, provide feedback and repair is useful for a dialog system. We show that in situated communication, eyetracking can be used to reliably and efficiently monitor the hearer's reference resolution process. An interactive system that draws on hearer gaze to provide positive or negative feedback after referring to objects outperforms baseline systems on metrics of referential success and user confusion.

1 Introduction

Because dialog is interactive, interlocutors are constantly engaged in a process of predicting and monitoring the effects of their utterances. Typically, a speaker produces an utterance with a specific communicative goal in mind—e.g., that the hearer will perform an action or adopt a certain belief—, and chooses one particular utterance because they *predict* that it will achieve this communicative goal. They will then *monitor* the hearer's reactions and infer from their observations whether the prediction actually came true. If they recognize that the hearer misunderstood the utterance, they may *repair* the problem by diagnosing what caused the misunderstanding and giving the hearer *feedback*. In a task-oriented dialog in which the hearer must perform a part of the task, feedback is especially important to inform the hearer when they made a mistake in the task. Ideally, the speaker should even detect when the hearer is *about to* make a mistake, and use feedback to keep them from making the mistake at all.

Many implemented dialog systems include a component for monitoring and repair. For instance, Traum (1994) presents a model for monitoring the grounding status of utterances in the TRAINS system; Young et al. (1994) show how the student's utterances in a dialog system can be used to uncover mistaken assumptions about their mental state; and Paek and Horvitz (1999) discuss an automated helpdesk system that can track grounding under uncertainty. However, most of these systems rely on the user's verbal utterances as their primary source of information; monitoring thus presupposes an (error-prone) language understanding module.

In the context of situated communication, where the speaker and hearer share a physical (or virtual) environment, one type of observation that can potentially give us a very direct handle on the hearer's understanding of an utterance is eye gaze. Eyetracking studies in psycholinguistics have shown that when listeners hear a referring expression, they tend to rapidly attend to the object in a scene to which they resolve this expression (Tanenhaus et al., 1995; Allopenna et al., 1998). For utterances that involve references to objects in the current environment, one can therefore ask whether eyetracking can be used to reliably judge the communicative success of the utterance. This would be of practical interest for implemented dialog systems once eyetracking becomes a mainstream technology; and even today, a system that reliably monitors communicative success using eyetracking could serve as a testbed for exploring monitoring and repair strategies.

In this paper, we present an interactive natural-language generation (NLG) system that uses eye-

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 30–39,
Seoul, South Korea, 5-6 July 2012.

tracking to monitor communicative success. Our system gives real-time instructions that are designed to help the user perform a treasure-hunt task in the virtual 3D environments of the recent Challenges on Generating Instructions in Virtual Environments (GIVE; Koller et al. (2010)). It monitors how the user resolves referring expressions (REs) by mapping the user's gaze to objects in the virtual environment. The system takes gaze to the intended referent as evidence of successful understanding, and gives the user positive feedback; by contrast, gaze to other objects triggers negative feedback. Crucially, this feedback comes before the user interacts with the object in the virtual environment, keeping the user from making mistakes before they happen.

We evaluate our system against one baseline that gives no feedback, and another that bases its feedback on monitoring the user's movements and their field of view. We find that the eyetracking-based system outperforms both on referential success, and that users interacting with it show significantly fewer signs of confusion about how to complete their task. This demonstrates that eyetracking can serve as a reliable source of evidence in monitoring communicative success. The system is, to our knowledge, the first dialog or NLG system that uses the hearer's gaze to monitor understanding of REs.

Plan of the paper. The paper is structured as follows. We first discuss related work in Section 2. We then describe our approach as well as the baselines in Section 3, set up the evaluation in Section 4 and present the results in Section 5. In Sections 6 and 7 we discuss our findings and conclude.

2 Related work

Dialog systems model a process of *grounding*, in which they decide to what extent the user has understood the utterance and the communicative goal has been reached. Observing the user behavior to monitor the state of understanding is a key component in this process. A full solution may require plan recognition or abductive or epistemic reasoning (see e.g. Young et al. (1994), Hirst et al. (1994)); in practice, many systems use more streamlined (Traum, 1994) or statistical methods (Paek and Horvitz, 1999). Most dialog systems focus on the verbal interaction of the system and user, and the user's utterances are therefore the primary source of evidence in the monitoring process. Some *incremental* dialog systems can monitor the user's verbal reactions to the system's utterances in real time, and continuously update the grounding state while the system utterance is still in progress (Skantze and Schlangen, 2009; Buss and Schlangen, 2010).

In this paper, we focus on the generation side of a dialog system—the user is the hearer—and on monitoring the user's *extralinguistic* reactions, in particular their gaze. Tanenhaus et al. (1995) and Allopenna et al. (1998) showed that subjects in psycholinguistic experiments who hear an RE visually attend to the object to which they resolve the RE. The "visual world" experimental paradigm exploits this by presenting objects on a computer screen and using an eyetracker to monitor the subject's gaze. This research uses gaze only as an experimental tool and not as part of an interactive dialog system, and the visual worlds are usually limited to static 2D scenes. Also, such setups cannot account for the reciprocal nature of dialog and the consequences that hearer gaze has for the speaker's monitoring process.

In the context of situated dialog systems, previous studies have employed robots and virtual agents as *speakers* to explore how and when speaker gaze helps human hearers to ground referring expressions (Foster, 2007). For instance, Staudte and Crocker (2011) show that an agent can make it easier for the (human) hearer to resolve a system-generated RE by looking at the intended referent, using head and eye movements. Conversely, the performance of a system for resolving human-produced REs can be improved by taking the (human) speaker's gaze into account (Iida et al., 2011). Gaze has also been used to track the general dynamics of a dialog, such as turn taking (Jokinen et al., in press).

Here we are interested in monitoring the *hearer's* gaze in order to determine whether they have understood an RE. To our knowledge, there has been no research on this; in particular, not in dynamic 3D environments. The closest earlier work of which we are aware comes from the context of the GIVE Challenge, a shared task for interactive, situated natural language generation systems. These systems typically approximate hearer gaze as visibility of objects on the screen and monitor grounding based on this (Denis, 2010; Racca et al., 2011).

Figure 1: A first-person view of a virtual 3D environment.

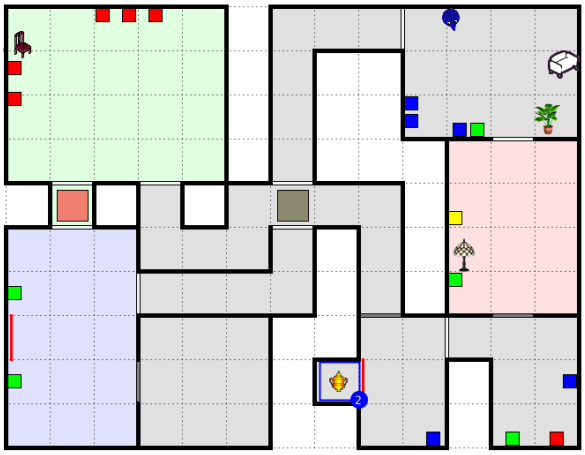

Figure 2: A map of the environment in Fig. 1; note the user in the upper right room.

3 Interactive natural-language generation in virtual environments

In this paper, we consider the communicative situation of the GIVE Challenge (Koller et al., 2010; Striegnitz et al., 2011). In this task, a human user can move about freely in a virtual indoor environment featuring several interconnected rooms and corridors. A 3D view of the environment is displayed on a computer screen as in Fig. 1, and the user can walk forward/backward and turn left/right, using the cursor keys. They can also press buttons attached to the walls, by clicking on them with the mouse once they are close enough. The small and big white circles in Fig. 1, which represent eyetracking information, are not actually visible to the user.

The user interacts with a real-time NLG system in the context of a treasure-hunt game, where their 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; however, they do not have prior knowledge of which buttons to press, so they rely on instructions and REs generated by the system. 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*. Next to buttons, rooms also contain a number of landmark objects, such as chairs and plants, which cannot directly be interacted with, but may be used in REs to nearby targets. Fig. 2 shows a top-down map of the virtual environment in which the scene of Fig. 1 arose. We call an entire game up to the successful discovery of the trophy, an *interaction* of the system and the user.

3.1 Monitoring communicative success

NLG systems in the GIVE setting are in an interactive communicative situation. This situation represents one complete half of a dialog situation: Only the system gets to use language, but the user moves and acts in response to the system's utterances. As a result, the system should continuously monitor and react to what the user does, in real time. This is most tangible in the system's use of REs. When a user misinterprets (or simply does not understand) a system-generated RE, there is a high chance that they will end up pressing the wrong button. This will hinder the completion of the task. A system that predicts how the user resolves the RE by monitoring their movements and actions, and that can proactively give the user feedback to keep them from making a mistake, will therefore perform better than one which cannot do this. Furthermore, if the system can give positive feedback when it detects that the user is about to do the right thing, this may increase the user's confidence.

Monitoring communicative success in GIVE interactions and providing the right feedback can be challenging. For example, in the original interaction from which we took the screenshot of Fig. 1, the system instructed the user to "push the right button to the right of the green button", referring to the rightmost blue button in the scene. In response, the user first walked hesitantly towards the far pair of buttons (green and blue), and then turned to face the other pair, as seen in Fig. 3. A typical NLG system used

Figure 3: The scene of Fig. 1, after the user moved and turned in response to a referring expression.

in the GIVE Challenge (e.g., Dionne et al. (2009), Denis (2010), Racca et al. (2011)) may try to predict how the user might resolve the RE based on the visibility of objects, timing data, or distances. Relying only on such data, however, even a human observer could have difficulties in interpreting the user's reaction; the user in Fig. 3 ended up closer to the green and blue buttons, but the other buttons (the two blue ones) are, to similar degrees, visually in focus.

The contribution of this paper is to present a method for monitoring the communicative success of an RE based on eyetracking. We start from the hypothesis that when the user resolves an RE to a certain object, they will tend to gaze at this object. In the scene of Fig. 3, the user was indeed looking at the system's intended referent, which they later pressed; the small white circles indicate a trace of recent fixations on the screen, and the big white circle marks the object in the virtual environment to which the system resolved these screen positions. Our system takes this gaze information, which is available in real time, as evidence for how the user has resolved its RE, and generates positive or negative feedback based on this.

3.2 NLG systems

To demonstrate the usefulness of the eyetracking-based approach, we implemented and compared three different NLG systems. All of these use an identical module for generating navigation instructions, which guides the user to a specific location, as well as object manipulation instructions such as "push the blue button"; "the blue button" is an RE that describes an object to the user. The systems generate REs that are optimized for being easy for the hearer to understand, according to a corpus-based model of understandability (Garoufi and Koller, 2011). The model was trained on human instructions produced in a subset of the virtual environments we use in this work. The resulting system computes referring expressions that are correct and uniquely describe the referent as seen by the hearer at the moment in which generation starts.

Unlike in the original GIVE Challenge, the generated instructions are converted to speech by the Mary text-to-speech system (Schröder and 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 and request a clarification. This will cause the system to generate an instruction newly; if it contains an RE, this RE may or may not be the same as the one used in the original utterance.

The difference between the three systems is in the way they monitor communicative success and determine when to give feedback to the user.

The no-feedback system. As a baseline system, we used a system which does not monitor success at all, and therefore never gives feedback on its own initiative. Notice that the system still re-generates an RE when the user presses the 'H' key.

Movement-based monitoring. As a second baseline, we implemented a system that attempts to monitor whether a user understood an RE based on their movements. This system is intended to represent the user monitoring that can be implemented, with a reasonable amount of effort, on the basis of immediately available information in the GIVE setting.

The movement-based system gives no feedback until only a single button in the current room is visible to the user, since it can be hard to make a reliable prediction if the user sees several buttons on their screen. 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 that is the button the system intended to refer to, the system utters

the positive feedback "yes, that one". For incorrect buttons, it utters the negative feedback "no, not that one". Although the negative feedback is relatively vague, it has the advantage of limiting the variability of the system's outputs, which facilitates evaluation.

Eyetracking-based monitoring. Finally, the eyetracking-based 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 determines the (x,y) position on the screen that the user is looking at. It then identifies the object in the environment that corresponds to this position by casting a ray from the (virtual) camera through the screen plane, and picking the closest object lying within a small range of this ray (Fig. 1; see Staudte et al. (2012) for details). If the user continously looks at the same object for more than a certain amount of time, the system counts this as an inspection of the object; for our experiments, we chose a threshold of 300 ms. Once the system detects an inspection to a button in the room, it generates positive or negative feedback utterances in exactly the same way as the movement system does.

Both the movement-based and the eyetracking-based model withhold their feedback until a first full description of the referent (a *first-mention RE*) has been spoken. Additionally, they only provide feedback once for every newly approached or inspected button and will not repeat this feedback unless the user has approached or inspected another button in the meantime. Example interactions of a user with each of the three systems are presented in Appendix A.

4 Evaluation

We set up a human evaluation study in order to assess the performance of the eyetracking system as compared against the two baselines on the situated instruction giving task. For this, we record participant interactions with the three systems employed in three different virtual environments. These environments were taken from Gargett et al. (2010); they vary as to the visual and spatial properties of the objects they contain. One of these environments is shown in Fig. 2. Overall, 31 participants (12 females) were tested. All reported their English skills as fluent, and all were capable of completing the tasks. Their mean age was 27.6 years.

4.1 Task and procedure

A faceLAB eyetracking system (`http://www.seeingmachines.com/product/facelab`) remotely monitored participants' eye movements on a 24-inch monitor, as in Fig. 4 and 5 of Appendix B. 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 any time they felt that the instructions were not sufficient (by pressing the 'H' key).

The eyetracker 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. Each participant started with a short practice session to familiarize themselves with the interface and to clarify remaining questions. We then collected three complete interactions, each with a different virtual environment and NLG system (alternated according to a Latin square design). Finally, each participant received a questionnaire which was aimed to reveal whether they noticed that they were eyetracked and that one of the generation systems made use of that, and how satisfied they were with this interaction. The entire experiment lasted approximately 30 minutes.

4.2 Analysis

For the assessment of communicative success in these interactions, we considered as referential scenes the parts of the interaction between the onset of a first-mention RE to a given referent and the participant's reaction (pressing a button or navigating away to another room). To control for external factors that could have an impact on this, 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

system	success			success w/out confusion			#scenes		
	all	easy	hard	all	easy	hard	all	easy	hard
eyetracking	93.4	100.0	90.4	91.9	100.0	88.2	198	62	136
with feedback	94.3	100.0	91.7	92.8	100.0	89.4	194	62	132
without feedback	50.0	-	50.0	50.0	-	50.0	4	0	4
no-feedback	86.6*	100.0°	80.6*	83.5**	98.9°	76.5**	284	88	196
movement	89.8°	100.0°	85.2°	87.5°	97.8°	82.8°	295	92	203
with feedback	93.9	100.0	90.6	91.9	97.7	88.7	247	88	159
without feedback	68.8	100.0	65.9	64.6	100.0	61.4	48	4	44

Table 1: Mean referential success rate (%) and number of scenes for the systems, broken down by scene complexity and presence of feedback. Differences of overall system performances to the eyetracking system are: significant at ** $p < 0.01$, * $p < 0.05$; ° not significant.

filtered out 9 interactions out of the 93 we collected.

Inferential statistics on this data were carried out using mixed-effect models from the lme4 package in R (Baayen et al., 2008). Specifically, we used logistic regression for modeling binary data, Poisson regression for count variables and linear regression for continuous data.

5 Results

On evaluating the post-task questionnaires, we did not find any significant preferences for a particular NLG system. Roughly the same number of them 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 participant mentioned the system's reaction to their eye gaze—though some noticed the (lack of) feedback. We take this to mean that the participants did not realize they were being eyetracked.

Below, we report results on objective metrics that do not depend on participants' judgments.

5.1 Confusion

A key goal of any RE generation system is that the user understands the REs easily. One measure of the ease of understanding is the frequency with which participants pressed the 'H' key to indicate their confusion and ask for help. The overall average of 'H' keystrokes per interaction was 1.14 for the eyetracking-based system, 1.77 for the movement-based system, and 2.26 for the no-feedback system. A model fitted to the keystroke 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, $p < .05$). In other words, the feedback given by the eyetracking-based system significantly reduces user confusion.

5.2 Referential success

An even more direct way to measure the interaction quality is the ratio of generated REs that the participants were able to resolve correctly. In our evaluation, we looked at two different definitions of success. First, an RE can count as successful if the first button that the user pressed after hearing the RE was the system's intended referent. The results of this evaluation are shown in the leftmost part of Table 1, under "success". A logistic mixed-effects model fitted to the referential success data revealed a marginal main effect of system ($\chi^2(2) = 5.55, p = .062$). Pairwise comparisons further show that the eyetracking system performs significantly better than the no-feedback system (Coeff. = -0.765, SE = 0.342, Wald's Z = $-2.24, p < .05$); no significant difference was found between the eyetracking-based and the movement-based system.

Second, we can additionally require that an RE only counts as successful if the user did not press the 'H' key between hearing the first-mention RE and pressing the correct button. This is a stricter version of referential success, which requires that the system recognized cases of potential confusion

and did not force the user to take the initiative in case of difficulties. It is in line with Dethlefs et al.'s (2010) findings that metrics that penalize difficulties the user encountered before successfully completing the task are better predictors of user satisfaction than ones that only consider the eventual task completion. Our results on this metric are shown in the middle part of Table 1, under "success without confusion". We observe again a main effect of system ($\chi^2(2) = 7.78, p < .05$); furthermore, the eyetracking system elicited again more correct buttons than the no-feedback system (Coeff. = -0.813, SE = 0.306, Wald's Z = $-2.66, p < 0.01$).

To obtain a more detailed view of when and to what extent the systems' behavior differed, we distinguished scenes according to their complexity. A scene was classified as *easy* if a) there were no distractors in it, or b) all distractors had different colors from the target, while the system included the color attribute in its RE. All other scenes were considered *hard*. Note that "easy" and "hard" are properties of the scene and not of the system, because every system generated the same REs in each scene.

In the experiments, we found essentially no difference between the success rates of different systems on easy scenes (see the "easy" columns of Table 1): All systems were almost always successful. The differences came almost exclusively from the hard scenes, where the eyetracking system performed significantly better than the no-feedback system (success: Coeff. = -0.793, SE = 0.348, Wald's Z = $-2.28, p < 0.05$; success without confusion: Coeff. = -0.833, SE = 0.315, Wald's Z = $-2.64, p < 0.01$) and, at least numerically, also much better than the movement system.

There was a particularly interesting difference in the feedback behavior of the eyetracking and movement systems on hard scenes (see the rightmost part of Table 1, labeled "#scenes"). In easy scenes, both systems almost always gave feedback (62/62 = 100.0%; 88/92 = 95.6%); but for hard scenes, the ratio of scenes in which the movement system gave feedback at all dropped to 159/203 = 78.3%, whereas the ratio for the eyetracking system remained high. This may have contributed to the overall performance difference between the two systems.

system	#actions (norm.)	distance (norm.)	duration (norm.)	idle (sec)
eyetracking	1.06	1.22	1.49	256.6
no-feedback	1.22*	1.27	1.59	272.5
movement	1.16	1.26	1.56	274.4

Table 2: Mean values of additional metrics. Differences to the eyetracking system are significant at * $p < 0.05$.

5.3 Further performance metrics

Finally, we measured a number of other objective metrics, including the number of actions (i.e., button presses), the distance the user traveled, the total duration of the interaction, and the mean time a participant spent idle. Even though these measures only partly provide statistically significant results, they help to draw a clearer picture of how the eyetracking-based feedback affects performance.

Because the three virtual environments were of different complexity, we normalized the number of actions, distance, and duration by dividing the value for a given interaction by the minimum value for all interactions of the same virtual environment. The resulting measures are shown in Table 2. Participants performed significantly fewer actions in the eyetracking system than in the no-feedback system (Coeff. = 0.174, SE = 0.067, t = $2.57, p(mcmc) < .05$); there were also trends that users of the eyetracking-based system traveled the shortest distance, needed the least overall time, and spent the least time idle.

The only measure deviating from this trend is movement speed, i.e., the speed at which users reacted to the systems' instructions to press certain buttons. For all successful scenes (without confusion), we computed the speed by dividing the GIVE distance (including turning distance) between the target referent and the user's location at the time of the instruction containing the first-mention RE by the time (in seconds) between hearing the instruction and pressing the target. The mean movement speed is 0.518 for the no-feedback system, 0.493 for the movement system, and 0.472 for the eyetracking system. A marginal main effect of movement speed confirms this trend ($\chi^2(2) = 5.58, p = .061$) and shows that participants moved more slowly when getting eyetracking-based feedback than when getting no feedback at all (Coeff. = 0.0352, SE =

0.0166, t = $-4.97, p(mcmc) < .05$).

6 Discussion

The results in Section 5 demonstrate the usefulness of eyetracking as a foundation for monitoring and feedback. Compared to the no-feedback system, the eyetracking-based system achieved a significantly lower confusion rate and a significantly higher RE success rate, especially on hard instances. The difference increases further if we discount scenes in which the user had to ask for help, thus forcing the system to give feedback anyway. In other words, eyetracking provides reliable and direct access to the hearer's reference resolution process. Real-time dialog systems can use gaze information to monitor the success of REs and generate feedback before the user actually makes a mistake.

Monitoring and feedback could also be achieved without using eyetracking. To explore this alternative, we compared eyetracking against a movement-based system. We found that the former outperformed the latter on hearer confusion and (at least numerically) on referential success, while not performing worse on other measures. This means that the improvement comes not merely from the fact that feedback was given; it is also important when and where feedback is given. The crucial weakness of the movement-based system is that it gave feedback for hard instances much more rarely than the eyetracking system. Increasing recall by lowering the system's confidence threshold would introduce fresh errors. Further improvements must therefore come at the cost of a more complex monitoring system, both conceptually and in terms of implementation effort. From this perspective, eyetracking offers good performance at low implementation cost.

One result that seems to go against the trend is that users of the eyetracking system moved significantly more slowly on their way to a target. We see two possible explanations for this. First, it may be that users needed some time to listen to the feedback, or were encouraged by it to look at more objects. A second explanation is that this is not really a difference in the quality of the systems' behavior, but a difference in the populations over which the mean speed was computed: The speed was only averaged over scenes in which the users resolved the RE correctly, and the eyetracking system achieved communicative success in many cases in which the others did not—presumably complex scenes in which the user had to work harder to find the correct button. This issue bears more careful analysis.

Finally, the eyetracking-based system could be improved further in many ways. On the one hand, it suffers from the fact that all objects in the 3D environment shift on the screen when the user turns or moves. 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 noise of this kind from the eyetracker signal. On the other hand, the negative feedback our system gave ("no, not that one") was quite unspecific. More specific feedback ("no, the BLUE button") might further improve the system's performance.

7 Conclusion

We described an interactive NLG system that uses eyetracking to monitor the communicative success of the REs it generates. The communication is situated in a virtual 3D environment in which the user can move freely, and our system automatically maps eyetracking screen coordinates to objects in the environment. A task-based evaluation found that the eyetracking-based system outperforms both a no-feedback system and a system whose feedback is based on the user's movements in the virtual environment, along with their field of view.

Eyetracking is currently widely available in research institutions, which should make our system easy to reimplement in other situated domains. We anticipate that eyetracking may become mainstream technology in the not-too-distant future. But even in a purely research context, we believe that the directness with which eyetracking allows us to observe the hearer's interpretation process may be useful as a testbed for efficient theories of grounding.

Acknowledgments. This research was partly supported by the Cluster of Excellence "Multimodal Computing and Interaction" at Saarland University. We are grateful to Irena Dotcheva for help with data collection as well as to Alexandre Denis and Christoph Clodo for software support, and to Kristina Jokinen for helpful comments.

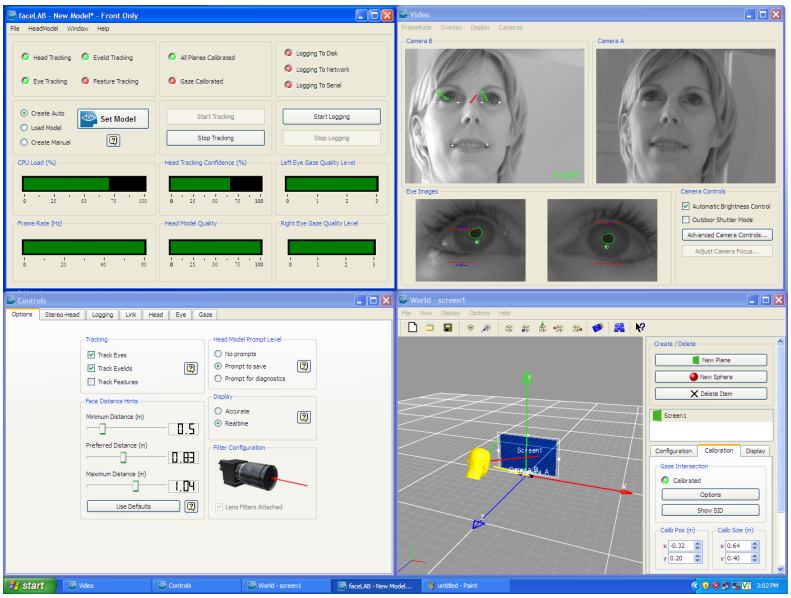

Figure 4: A screenshot from the faceLAB software, including visualization of eye-gaze position in 3D space.

References

Paul Allopenna, James Magnuson, and Michael Tanenhaus. 1998. Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language*, 38:419–439.

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

Okko Buss and David Schlangen. 2010. Modelling sub-utterance phenomena in spoken dialogue systems. In *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 33–41.

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

Nina Dethlefs, Heriberto Cuayahuitl, Kai-Florian Richter, Elena Andonova, and John Bateman. 2010. Evaluating task success in a dialogue system for indoor navigation. In *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 143–146.

Daniel Dionne, Salvador de la Puente, Carlos León, Pablo Gervás, and Raquel Hervás. 2009. A model for human readable instruction generation using level-based discourse planning and dynamic inference of attributes. In *Proceedings of the 12th European Workshop on Natural Language Generation*.

Mary Ellen Foster. 2007. Enhancing human-computer interaction with embodied conversational agents. In *Proceedings of HCI International 2007*.

Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Kristina Striegnitz. 2010. The GIVE-2 Corpus of Giving Instructions in Virtual Environments. In *Proceedings of the 7th Conference on International Language Resources and Evaluation*.

Konstantina Garoufi and Alexander Koller. 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*.

Graeme Hirst, Susan McRoy, Peter Heeman, Philip Edmonds, and Diane Horton. 1994. Repairing conversational misunderstandings and non-understandings. *Speech Communications*, 15:213–229.

Ryu Iida, Masaaki Yasuhara, and Takenobu Tokunaga. 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*.

K. Jokinen, H. Furukawa, M. Nishida, and S. Yamamoto. in press. Gaze and turn-taking behaviour in casual conversational interactions. *ACM Trans. Interactive Intelligent Systems*. Special Issue on Eye Gaze in Intelligent Human-Machine Interaction.

Alexander Koller, Kristina Striegnitz, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, and Jon

Oberlander. 2010. The First Challenge on Generating Instructions in Virtual Environments. In Emiel Krahmer and Mariet Theune, editors, *Empirical Methods in Natural Language Generation*, number 5790 in LNCS, pages 337–361. Springer.

Tim Paek and Eric Horvitz. 1999. Uncertainty, utility, and misunderstanding: A decision-theoretic perspective on grounding in conversational systems. In *AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*.

David Nicolás Racca, Luciana Benotti, and Pablo Duboue. 2011. The GIVE-2.5 C Generation System. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*.

Marc Schröder and J. Trouvain. 2003. The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology*, 6:365–377.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*.

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

Maria Staudte, Alexander Koller, Konstantina Garoufi, and Matthew W. Crocker. 2012. Using listener gaze to augment speech generation in a virtual 3D environment. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. To appear.

Kristina Striegnitz, Alexandre Denis, Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Mariet Theune. 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*.

Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.

David Traum. 1994. *A computational theory of grounding in natural language conversation*. Ph.D. thesis, University of Rochester.

Michael Young, Johanna Moore, and Martha Pollack. 1994. Towards a principled representation for discourse plans. In *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society*.

A Example interactions

The following interactions between a user (U) and each of the three systems (S) were recorded during the systems' attempts to instruct the user to press the rightmost blue button shown in Fig. 1.

A.1 Eyetracking system

(1) S: *Push the right button to the right of the green button.*
U: (approaches the pair of blue and green button and inspects one of them)
S: *No, not that one!*
. . . (U inspects other buttons in the scene, while S provides appropriate feedback)
U: (inspects the correct target)
S: *Yes, that one!*
U: (presses the correct button)

A.2 Movement system

(2) S: *Push the right button to the right of the green button.*
U: (approaches the pair of blue and green buttons; once the user is very close to the blue button, it happens to become the only button visible on screen)
U: (continues moving closer to the blue button)
S: *No, not that one!*
U: (has no time to react to the system's feedback and presses the wrong blue button)

A.3 No-feedback system

(3) S: *Push the right button to the right of the green button.*
U: (presses the wrong blue button)

B The experimental setup

Figure 5: A faceLAB eyetracking system monitored participants' eye movements during the interactions.