Using Listener Gaze to Refer in Installments Benefits Understanding

Nikolina Mitev[1] **(nikkol@coli.uni-saarland.de)**
Patrick Renner[2] **(patrick.renner@uni-bielefeld.de)**
Thies Pfeiffer[2] **(thies.pfeiffer@uni-bielefeld.de)**
Maria Staudte[1] **(masta@coli.uni-saarland.de)**
[1] Embodied Spoken Interaction Group, Saarland University, Saarbrücken, Germany
[2] Cluster of Excellence Cognitive Interaction Technology, Bielefeld University, Germany

Abstract

Listener gaze can predict reference resolution as it reflects listeners' understanding. Further, speakers commonly refer in installments to co-present objects by providing a description incrementally. Here, we investigate whether listener gaze could be utilized to refer incrementally, in spoken installments. Specifically, we implemented a system that generates instructions, describes objects, and reacts to listener gaze with verbal feedback. We compared unambiguous vs. ambiguous instructions supplemented by two levels of feedback specificity: either underspecified ("No, not that one!") or more informative, contrastive responses ("Further left!"). Our findings show that ambiguous instructions with underspecified feedback did not benefit task performance. In contrast, ambiguous instructions with contrastive feedback (referring in installments) resulted in more efficient interactions. Moreover, this strategy even outperformed the one providing unambiguous instructions.

Keywords: Human-Machine Interaction, Listener Gaze, Referring in Installments, Reference Resolution, Gaze-sensitive Feedback

Introduction

Reference resolution plays an important role in achieving communicative success in situated task-oriented interaction. A spoken utterance in natural language is usually not interpreted in isolation but other non-verbal channels are considered to reflect understanding (Clark & Krych, 2004; Hanna & Brennan, 2007; Brown-Schmidt, 2012). Specifically, listener gaze can reflect reference resolution because listeners look at those objects present in the visual context that they believe a speaker refers to (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In the present work, we further examine how an artificial speaker can make use of this in order to make instruction-giving more efficient.

In human-machine interaction, Garoufi, Staudte, Koller, and Crocker (2016) showed that exploiting listener gaze benefits interactions with a natural language generation (NLG) system. The NLG system gave instructions to guide a human listener through a virtual maze and to refer to specific buttons to be pressed. Additionally, the system provided verbal feedback on the basis of button inspections to confirm or reject these without providing any further information. And indeed, interacting with the gaze-sensitive system led to better performance than interacting with a baseline system without feedback. Koleva, Villalba, Staudte, and Koller (2015) carried out a follow-up study and investigated automatic prediction of success of referential resolution by the listener. They demonstrated on the collected corpus that eye tracking features combined with other observational features improve prediction especially in complex referential scenes where more competitors next to the target are available in the visual context.

In human-human interaction, Brennan, Schuhmann, and Batres (2013) considered outdoor navigation in a real environment and examined the kind of referring expressions and lexical choice during remote pedestrian guidance. They observed effects of spatial ability as well as a strong degree of entrainment but did not examine listener gaze. This was addressed by Koleva, Hoppe, Moniri, Staudte, and Bulling (2015), who manipulated gaze availability to the speaker during indoor guidance. The speaker monitored listener behavior by watching a scene video of the walker's egocentric perspective overlaid with either the true, a perturbed or no gaze cursor. They report no benefit of showing the gaze position to the human instructor suggesting that speakers are already extremely efficient at producing referring expressions in such a setup. Yet speakers produced more often negative feedback instances in the presence of the gaze cursor and in turn listeners showed deliberate use of gaze right before and just after instructions.

Further evidence from human-human interactions suggests that speakers commonly provide information in subsequent chunks to the listener. Referring to objects in such chunks has been termed *referring in installments* (Striegnitz, Buschmeier, & Kopp, 2012). Installments facilitate the adaptation to changes in the visual context and also to the listener's signals. Zarrieß and Schlangen (2016) demonstrated that an NLG system can make use of such an interaction strategy and lead to a higher task performance for object identification on static images. However, currently there is no evidence if listener gaze can be used to apply installments interactively and how effective such an approach would be for the generation of referring expressions.

In this paper we address this issue. We designed and implemented an interactive NLG system that describes real-world objects in complex scenes to a human listener in an assembly scenario. The system monitors listener gaze and uses this in real-time to provide verbal feedback to object inspections. We compared two interaction strategies and tested their effectiveness: a) generating a long, UNAMBIGUOUS instruction vs. b) generating a short, AMBIGUOUS instruction followed by gaze-driven feedback. Furthermore, we examined the impact of feedback specificity by either providing an UNDERSPECIFIED or a CONTRASTIVE response expressing the spatial relation of the target relative to the 3D position cur-

rently gazed at. The aim was to assess whether and how the more informative responses affect the interaction and whether listeners can actually benefit from the more interactive and piece-wise delivery of information based on listener gaze.

Our work is related to the work of Eaddy, Blasko, Babcock, and Feiner (2004) who employed an assistance system and used gaze to tune verbal feedback. Yet no systematic evaluation is reported and also to what extent the feedback was automatically generated remains unclear. Their and our proposed system realize an interaction design of a perceptual user interface as previously suggested by Turk and Robertson (2000). Our results connect well to previous findings from work with virtual environments, namely that an NLG system is more helpful when providing gaze-driven feedback. Moreover, we show that AMBIGUOUS instructions followed by CONTRASTIVE feedback (referring in form of installments) led to more efficient interactions than when they were followed by UNDERSPECIFIED feedback. Surprisingly, the former even outperformed the UNAMBIGUOUS interaction strategy.

GazInG: An interactive NLG system

GazInG is an interactive instruction-giving system that monitors listener gaze and generates proactive spoken feedback on the basis of object inspections. The system is flexible, extendable and adoptable to other domains due to its modular design (Figure 1). The `InteractionManager` is the core software component. It steers the interaction flow and is coupled to the `EyeSee3D` module (Pfeiffer & Renner, 2014) that transfers the real scene into a virtual 3D situation model and realizes the semantic mapping of object inspections. The `InteractionManager` has access to the `Domain` knowledge, where the characteristics of the real-world objects are stored. The auditory instruction was generated by the `SpeechSynthesizer` MaryTTS (Schröder, Charfuelan, Pammi, & Steiner, 2011). The programming language used for the implementation of the NLG system is Java, EyeSee3D is implemented in C++ and JavaScript. The synchronization of the different modalities is necessary to realize a smooth situated interaction. They are aligned using multi-threading and event-based programming. The inspection threshold was initially set to 300 ms inspired by Garoufi et al. (2016), who dealt with long distances between the listener and the target. However, we adjusted it to 200 ms as we are dealing with short distances and targets close at hand.

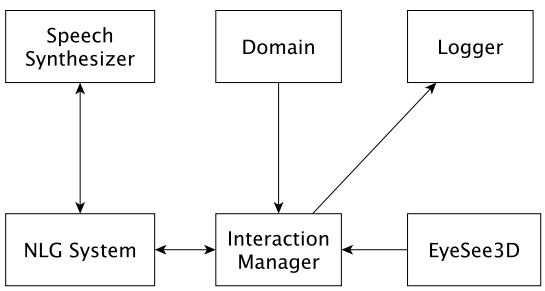

Figure 1: This diagram depicts the modular software architecture of the GazInG system.

Natural Language Generation

NLG heuristics Each object in the workspace (described in **Setup and Apparatus**) is composed of two simple building blocks. We applied a heuristic approach to generate an identifying instruction containing a referring expression that describes a composed object. The syntactic structure of the instructions is fixed. An AMBIGUOUS instruction consists of a main clause describing the bottom object (1). The head noun is randomly chosen from a set of synonyms suitable for the type of object. Size and color are included as premodifiers. For the generation of an UNAMBIGUOUS instruction the system extends an AMBIGUOUS instruction with two post-modifiers: a) a prepositional phrase (PP) or a relative clause (RelClause) to specify the characteristics of the top object and b) an adverbial phrase expressing the absolute viewer-centered position of the target (2). The workspace is divided into four squares at the back towards the left or the right and in the front towards the left or the right.

(1) AMBIGUOUS "Pick the big red building block!"

(2) UNAMBIGUOUS "Pick the big red building block [with the small yellow one on top]$^{a)}$ [at the back towards the left]$^{b)}$!"

Verbal Feedback The system generates either UNDERSPECIFIED or CONTRASTIVE feedback. Gazing at a target object triggers positive feedback (e.g. *"Yes"*, *"Exactly"* etc.). The consideration of competitor objects elicits negative feedback, either UNDERSPECIFIED (e.g. *"No, not that one!"*) or CONTRASTIVE providing relative direction information (e.g. *"Further left!"*). While UNDERSPECIFIED feedback excludes only the currently inspected competitor, which may be sufficient for simple scenes with fewer competitors, CONTRASTIVE feedback not only discards an intention but directs listeners' attention towards the target from her current gaze position. In this way the system avoids inspections of other competitors until finding the target and implements the notion of referring in installments.

Experiment

In our experiment we used a mixed factorial design. The *InteractionStrategy* chosen by the GazInG system was a within subjects independent variable, i.e. every participant experienced UNAMBIGUOUS and AMBIGUOUS instructions. In contrast, *FeedbackSpecificity* was manipulated between subjects.

Participants Forty-eight participants, mainly students enrolled at Saarland University, took part in the experiment. The average age of the first group of participants was 25 years with a range of 19–35 and in the second 24 years with a

	Interaction Strategy	
	AMBIGUOUS	UNAMBIGUOUS
GROUP 1	Underspecified Feedback	No Feedback
GROUP 2	Contrastive Feedback	Contrastive Feedback

Table 1: Interaction strategies (blocked) for each group.

range of 20–31. All participants were German native speakers and reported normal or corrected-to-normal vision and no red-green color blindness. Their participation was compensated with €8 (first group) and €5 (second group) due to the slightly shorter duration of the experiment.

Procedure Participants were seated in front of the workspace and were provided with the task description. They were told to follow the system's spoken instructions on which object to grasp next and to team up with the GazInG system and solve the task together as precisely as possible. After that participants were equipped with a pair of eye tracking glasses and a 3-point calibration procedure was carried out. Calibration was repeated between layouts and whenever necessary. Prior to the experimental part, participants completed a short practice session in order to familiarize with the task and the system's pace. They had to collect three targets among six objects in total. The experimental part consisted of two blocks, one for each *InteractionStrategy* (see Table 1). The order was balanced across participants. In each block (made up of one layout) eight targets among 20 composed objects in total had to be identified and selected. The GazInG system did not instruct the listener on how to assemble the building blocks. However, participants were encouraged to be creative by giving an additional reward to the most creative LEGO model. Subsequent to finishing one layout, participants filled in a questionnaire answering questions about their perception and impressions of the interaction they just experienced. At the end, they answered questions about the comparison of both strategies. The duration of the experiment was between 30 and 45 minutes.

Setup and Apparatus

On Figure 2 our setup is depicted. The domain we chose for our scenario is LEGO DUPLO. It is appropriate for our setup because the building blocks are of convenient size, allow various combinations and different ways of assembly. We made use of SMI Eye Tracking Glasses, a binocular head-mounted eye tracker, to obtain gaze data. The tracker has a high resolution scene camera (1280 x 960) at 24Hz and two eye cameras recording at 30Hz. The glasses were connected to a notebook on which the EyeSee3D augmented reality software and the NLG system were running. We used a Dell Precision M4800 15,6" WORKSTATION with processor I7 4900MQ at 2.8GHZ and with 16GB RAM. The speech synthesizer was located on a remote server and accessed by a client-server architecture.

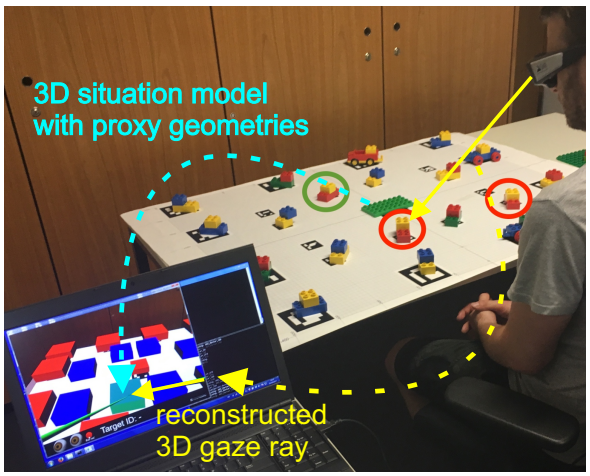

Figure 2: Setup: Listener in front of a workspace before any objects are collected. The target is circled in green and competitors in red. The listener inspects the competitor object to the left as highlighted in the virtual 3D model. EyeSee3D is used to reconstruct the gaze ray in 3D (yellow). The target domain is modeled as a 3D situation model with boxes as proxies for the assembled structures (turquoise).

Analysis

All measures were collected on a per item basis. The overall performance measure indicating efficiency is the task completion time. Figure 3 illustrates the three phases of the interaction: The duration of the spoken **initial instruction**, from speech onset to offset. The **search** phase being the time interval from instruction offset to the first inspection of the correct target, which we analyzed by evaluating the eye tracking data. Though for the UNAMBIGUOUS interaction strategy there is an overlap and the **search** is expected to be shorted because it begins already during listening contrary to the AMBIGUOUS one. The last phase is the time until a target was **grasp**ed.

The language modality was the independent variable and manipulated throughout the experiment. However, the verbal feedback varies because it is driven by the gaze behavior of the listener. Thus we analyzed the number of feedback instances and also the sequential order they occurred, i.e. time intervals from instruction offset to feedback onset of the first positive and also first negative feedback instance.

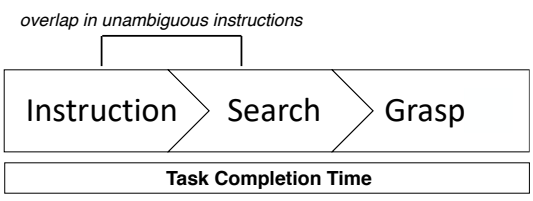

Figure 3: For the analysis the interaction was divided in three phases.

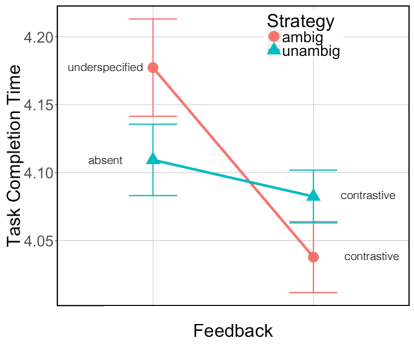

Figure 4: Task completion time from the instution onset until the target is grasped (log transformed with 95% CI - error bars).

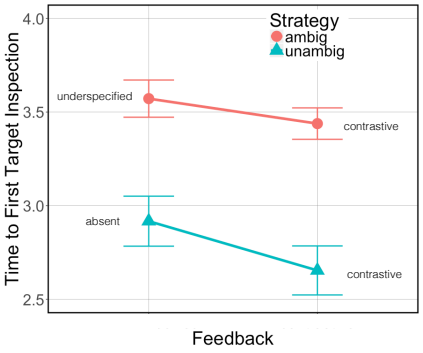

Figure 5: Time span from the instruction offset to the first target inspection (log transformed with 95% CI - error bars) .

For the statistical analyses we used the R statistical programming environment (R Core Team, 2014). We ran linear mixed-effects models using the lme4 package in R (Barr, Levy, Scheepers, & Tily, 2013) and model comparison in order to determine the influence of *InteractionStrategy* and *FeedbackSpecificity*. All effects were further validated by applying a mixed-design ANOVA (F1 as well as F2 analysis) on this dataset using the ez package (Lawrence, 2011).

Results

Performance: Task Completion Time

Task completion time indicates efficiency of the communication with the GazInG system. We obtain very high success rates as only few wrong grasps were detected (8.7%). Our main findings are visualized in Figure 4. Providing CONTRASTIVE feedback after an UNAMBIGUOUS instruction was beneficial and participants were faster compared to when feedback was absent (blue line). Moreover, CONTRASTIVE feedback led to better performance in the AMBIGUOUS condition than UNDERSPECIFIED feedback and, notably, even outperforms the UNAMBIGUOUS interaction strategy (red line). Specifically, there was a main effect of *FeedbackSpecificity* on task completion time revealed by model comparison ($\chi^2(1) = 10.513, p < .01$). Further, a significant interaction of *InteractionStrategy* and *FeedbackSpecificity*, ($\chi^2(1) = 19.038, p < .001$) was found. Group one solved the task faster under UNAMBIGUOUS instruction without feedback ($M = 14307.44\,ms, SD = 8597.188\,ms$) than under AMBIGUOUS instruction with UNDERSPECIFIED feedback ($M = 17557.62\,ms, SD = 10441.538\,ms$). Group two received CONTRASTIVE feedback in both conditions and the effect changed its direction, the AMBIGUOUS condition now led to shorter task completion time ($M = 11955.16\,ms, SD = 5605.423\,ms$) compared to the UNAMBIGUOUS one ($M = 12745.13\,ms, SD = 4745.216\,ms$).

Listener Gaze: Visual Search Behavior

Further, we took a closer look to the time interval from instruction offset until finding and inspecting a target, i.e. search phase (see Figure 5). As expected, this interval was shorter in the UNAMBIGUOUS condition because the instructions specify all target characteristics including its position and so the search takes place as an instruction is unfolding. Analogously to the analysis of the performance measures above, we fitted linear mixed-effects models to this data set and observed main effects of *InteractionStrategy* ($\chi^2(1) = 36.390, p < .001$) and *FeedbackSpecificity* ($\chi^2(1) = 7.386, p < .01$). That is, listeners searched three times longer following an AMBIGUOUS instruction with UNDERSPECIFIED feedback ($M = 7219.720\,ms, SD = 8367.127\,ms$) than following an UNAMBIGUOUS one ($M = 2166.006\,ms, SD = 5122.780\,ms$). Gaze-driven CONTRASTIVE feedback also significantly shortened this interval. Again, in the UNAMBIGUOUS interaction strategy, listeners inspected the intended target sooner ($M = 1264.525\,ms, SD = 2207.747\,ms$) than in the AMBIGUOUS one ($M = 4213.038\,ms, SD = 3800.032\,ms$).

Table 2 summarizes the trial proceeding consisting of 1) listening to an instruction, 2) visual search, i.e. the time to first target inspection, and 3) the time until the target is grasped.

Language: Feedback Occurrences

Since group one did not receive any feedback in the UNAMBIGUOUS interaction strategy, there is no comparison between groups for that condition. We analyzed the number of negative feedback instances which occurred in the AMBIGUOUS condition across groups. To test if there is a significant difference, we constructed a generalised linear mixed-effects model (with a logit link function) fitted to *FeedbackOccur-*

		Initial Instruction	Search	Grasp	Total
GROUP 1	UNAMB.	7213.311	2166.006	4928.122	14307.44
	AMB.	2812.840	7219.720	7525.057	17557.62
GROUP 2	UNAMB.	7231.044	1264.525	4249.557	12745.13
	AMB.	2800.701	4213.038	4941.424	11955.16

Table 2: A trial consists of three phases (see Figure 3). Here the mean durations in milliseconds are given.

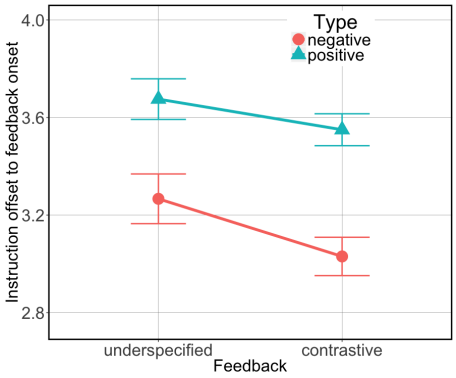

Figure 6: Time interval from the instruction offset to the onsets of the first negative and first positive feedback instances for the AMBIGUOUS *InteractionStrategy* (log transformed with 95% CI - error bars).

rences with *FeedbackSpecificity* as a fixed effect. Surprisingly, there was no significant difference ($\beta = -0.038$, $SE = 0.086$, $z = -0.443$, $p = .658$). Overall there were more positive than negative instances ($\beta = -0.094$, $SE = 0.048$, $z = -1.948$, $p = .0514$). Once the target is identified by the listener, she continues fixating it until it is reached which triggers positive feedback. Negative feedback instances could occur after a positive feedback instance because listeners turn quickly to assemble the target. Additionally no neutral fixation position exists such as a fixation cross in the visual world paradigm. For this reason, we carried out sequential analysis for the feedback occurrences, i.e. how long after instruction offset listeners heard negative and positive feedback for the first time. These feedback instances were triggered by the first inspection of a relevant competitor and of the target respectively and mirror visual search behavior. Figure 6 depicts the mean time intervals from the instruction offset to the feedback onset for the AMBIGUOUS condition. In line with the expected interaction sequences, positive feedback appeared later than negative feedback, revealed by a main effect of *FeedbackType* ($\chi^2(1) = 123.455, p < .001$). More importantly, this time with *FeedbackPresence* as a factor there was a main effect of *FeedbackSpecificity* ($\chi^2(1) = 18.416, p < .001$). As expected, for the time to the first positive feedback instance we observed the same pattern as for the time to the first fixation. Positive feedback occurred later in the interactions when feedback was UNDERSPECIFIED ($M = 10328.157ms, SD = 16912.019ms$) compared to when it was CONTRASTIVE ($M = 5427.947ms, SD = 5966.655ms$). This observation indicates that more informative feedback successfully reduces the search space and the time until identifying the target. Moreover, the informativity of the feedback similarly influences the time until a competitor fitting the description is inspected. This was revealed by the investigation of the first negative feedback occurrence: faster competitor consideration under CONTRASTIVE ($M = 1972.350ms, SD = 2683.437ms$) than under UNDERSPECIFIED ($M = 4067.041ms, SD = 5767.473ms$) feedback.

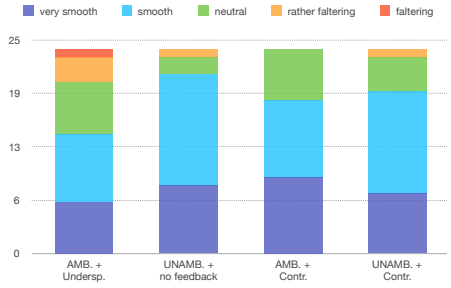

Figure 7: Participants' judgement of the interaction flow measured on a Likert scale in all four conditions.

This demonstrates that whenever the listener is expecting a more informative system response, she uses her gaze more deliberately in order to get an useful message and be able to progress in the task.

Perception: Questionnaires

Independent of *InteractionStrategy*, the interactions with the GazInG system were judged as natural and participants would be ready to use such a system in their daily life when assembling something. Listeners rated gaze-driven feedback to be helpful and not confusing. Altogether the system was well perceived in terms of pace and flow, which validates our design and choice of system parameters. Further, we assessed which *InteractionStrategy* they preferred. Interestingly, both groups voted for the one providing UNAMBIGUOUS instruction, even though participants in the second group, who experienced the AMBIGUOUS instructions with CONTRASTIVE feedback (referring in installments strategy), were more efficient. However, the interaction flow for AMBIGUOUS instructions with CONTRASTIVE feedback was judged to be better than with UNDERSPECIFIED feedback (as illustrated on Figure 7). The former was rated as smooth or very smooth by 75% (third bar) as opposed to the latter only 58% (first bar). The interaction flow for the UNAMBIGUOUS strategy followed by CONTRASTIVE feedback was similar AMBIGUOUS instruction followed by CONTRASTIVE feedback (rated as smooth or very smooth by 79% of the participants (fourth bar)). This shows that distributing the information in installments (partial instruction with gaze-driven feedback) is perceived as smooth as to follow UNAMBIGUOUS instructions.

Discussion

There is some evidence that human instruction givers may not benefit from the availability of listener gaze, at least in such a specific setting, with listener gaze projected as a cursor (Koleva, Hoppe, et al., 2015). In contrast, an artificial speaker (NLG system) can effectively exploit listener gaze for feedback generation in real environments. Thus, our results are in line with the previous finding from virtual environments (Garoufi et al., 2016), showing that providing verbal feedback after UNAMBIGUOUS instructions shortened interaction time compared to when feedback was absent.

Furthermore, we provide evidence that an NLG system can use listener gaze to refer to objects efficiently, in installments, by increasing the informativity of the gaze-driven feedback. That is, CONTRASTIVE feedback shortened task completion time and visual search. After an AMBIGUOUS instruction, listeners inspected the target object sooner with CONTRASTIVE than with UNDERSPECIFIED feedback. This indicates that the former eliminated wrong intentions faster and, thus, narrowed down the search for the target. Interestingly, the first inspection of a competitor object also occurred sooner when feedback was CONTRASTIVE. We interpret this to reflect a more deliberate use of gaze by the listener, driven by the expectation of eliciting a more informative response from the system.

It could further be argued that giving an AMBIGUOUS instruction is rather unnatural and awkward, especially in overloaded scenes with many competitors. However, providing the listener with supplementary information in form of proactive gaze-driven feedback turns out to be even more efficient than following a possibly more natural, UNAMBIGUOUS instructions (significant interaction of *InteractionStrategy* and *FeedbackSpecificity*). It is possible that following a long and exhaustive instruction is more demanding due to memorizing the many details of the object description at once. However, the perception results revealed that participants still preferred UNAMBIGUOUS instructions. We interpret this dichotomy to indicate that listeners were simply more self confident in the UNAMBIGUOUS condition about their choices of target objects. Additionally, with AMBIGUOUS instructions, listeners had to *actively engage* with the system in order to progress the interaction and identify the target. The UNAMBIGUOUS instruction, in contrast, allowed them to be more passive and wait until all relevant information was gathered. Such a behavior may also contribute to perceiving the UNAMBIGUOUS condition as more convenient – despite being less efficient.

Conclusion

Our work provides evidence that listeners (in Human-Computer Interaction) benefit from an interactive and piecewise delivery of information, namely in form of installments consisting of a partial instruction combined with feedback based on listener gaze. Specifically, we implemented an interactive NLG system and conducted an experiment to investigate the usefulness of gaze-driven feedback after (non-)exhaustive descriptions. In this experiment, we examined UNAMBIGUOUS vs. AMBIGUOUS instructions: the former supplemented with no or CONTRASTIVE, and the latter supplemented with UNDERSPECIFIED or CONTRASTIVE feedback. Our results are related to previous research and further reveal that AMBIGUOUS instructions with UNDERSPECIFIED feedback were less efficient than UNAMBIGUOUS instructions. However, the combination of AMBIGUOUS instruction with CONTRASTIVE feedback (referring in installments) was significantly more efficient and, in fact, even outperformed the UNAMBIGUOUS interaction strategy providing all the information needed to identify a target at once.

Acknowledgments

This work was funded by the "Multimodal Computing and Interaction" Cluster of Excellence at Saarland University and in parts by the Cluster of Excellence Cognitive Interaction Technology CITEC (EXC 277) at Bielefeld University.

References

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Brennan, S. E., Schuhmann, K. S., & Batres, K. M. (2013). Entrainment on the move and in the lab: The walking around corpus. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society.* Berlin, Germany.

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, *27*(1), 62–89.

Clark, H. H., & Krych, M. A. (2004, January). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*(1), 62–81. doi: 10.1016/j.jml.2003.08.004

Eaddy, M., Blasko, G., Babcock, J., & Feiner, S. (2004). My own private kiosk: Privacy-preserving public displays. In *Eighth International Symposium on Wearable Computers, 2004. ISWC 2004.* (Vol. 1, pp. 132–135).

Garoufi, K., Staudte, M., Koller, A., & Crocker, M. W. (2016). Exploiting listener gaze to improve situated communication in dynamic virtual environments. *Cognitive Science*, *40*(7), 1671–1703.

Hanna, J. E., & Brennan, S. E. (2007, November). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*(4), 596–615.

Koleva, N., Hoppe, S., Moniri, M. M., Staudte, M., & Bulling, A. (2015). On the interplay between spontaneous spoken instructions and human visual behaviour in an indoor guidance task. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015, Pasadena, California, USA, July 22-25, 2015.*

Koleva, N., Villalba, M., Staudte, M., & Koller, A. (2015). The impact of listener gaze on predicting reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers* (pp. 812–817).

Lawrence, M. A. (2011, 01). ez: Easy analysis and visualization of factorial experiments.

Pfeiffer, T., & Renner, P. (2014). EyeSee3D: a low-cost approach for analyzing mobile 3d eye tracking data using computer vision and augmented reality technology. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 369–376).

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Schröder, M., Charfuelan, M., Pammi, S., & Steiner, I. (2011, 8). Open source voice creation toolkit for the MARY TTS Platform. In *Proceedings of Interspeech 2011.* ISCA.

Striegnitz, K., Buschmeier, H., & Kopp, S. (2012). Referring in installments: A corpus study of spoken object references in an interactive virtual environment. In *Proceedings of the 7th International Natural Language Generation Conference* (pp. 12–16).

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Turk, M., & Robertson, G. (2000, March). Perceptual user interfaces (introduction). *Commun. ACM*, *43*(3), 32–34.

Zarrieß, S., & Schlangen, D. (2016). Easy things first: Installments improve referring expression generation for objects in photographs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016).*