On the interplay between spontaneous spoken instructions and human visual behaviour in an indoor guidance task

Nikolina Koleva (nikkol@coli.uni-saarland.de)
Embodied Spoken Interaction Group, Saarland University, Saarbrücken, Germany

Sabrina Hoppe (shoppe@mpi-inf.mpg.de)
Perceptual User Interfaces Group, Max Planck Institute for Informatics, Saarbrücken, Germany

Mohammad Mehdi Moniri (Mohammad_Mehdi.Moniri@dfki.de)
German Research Center for Artificial Intelligence, Saarbrücken, Germany

Maria Staudte (masta@coli.uni-saarland.de)
Embodied Spoken Interaction Group, Saarland University, Saarbrücken, Germany

Andreas Bulling (bulling@mpi-inf.mpg.de)
Perceptual User Interfaces Group, Max Planck Institute for Informatics, Saarbrücken, Germany

Abstract

We present an indoor guidance study to explore the interplay between spoken instructions and listeners' eye movements. The study involves a remote speaker to verbally guide a listener and together they solved nine tasks. We collected a multi-modal dataset consisting of the videos from the listeners' perspective, their gaze data, and instructors' utterances. We analyse the changes in instructions and listener gaze when the speaker can see 1) only the video, 2) the video and the gaze cursor, or 3) the video and manipulated gaze cursor. Our results show that listener visual behaviour mainly depends on utterance presence but also varies significantly before and after instructions. Additionally, more negative feedback occurred in 2). While piloting a new experimental setup, our results provide indication for gaze reflecting both: a symptom of language comprehension and a signal that listeners employ when it appears useful and which therefore adapts to our manipulation.

Keywords: referential situated communication; specific task guidance; mobile eye tracking; visual behaviour analysis; gaze-sensitive feedback

Introduction

We constantly direct our gaze to different parts of the visual scene to be able to perceive objects of interest with high acuity. These eye movements can be driven internally, i.e. by some self-initiated goal or intent, or externally by something that attracts our visual attention (Yantis & Jonides, 1990). External factors driving the attention can be salient objects in the visual scene or another person's utterances that direct our eyes to a co-present object or event. The latter has been exploited in many psycholinguistic studies in order to study language comprehension processes (for example see Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995).

Conversely, a listener's gaze may also signal (mis-) understanding back to the speaker. Taking the listener's behaviour into account when planning and making utterances is an important aspect of collaborative, goal-oriented interaction. In this sense, a listener's eye movements can be both: A result of a comprehension process, i.e. a "symptom", and/or a "signal" and feedback channel to the speaker, who can then modify and adapt their next utterance. A listener may even consciously use her gaze, similar to a pointing gesture, for instance in order to point when her hands are full or for other reasons unavailable.

This reciprocal nature of gaze during spoken interactions is not captured in most interactive studies so far, also because it is difficult to assess. Eye movements may be considered both a dependent variable (*symptom*, as an indicator for comprehension processes) and an indirect independent variable (*signal*, affecting utterance content). The aim of the present study is to shed light onto this dual role of gaze and to quantify to which extent listener eye movements depend on the speaker's utterance and vice versa.

We designed an exploratory experiment that involves spontaneous spoken instructions in a real-world environment while we manipulated the availability of listener gaze (henceforth *GazeAvailability*) in form of a cursor to the speaker. Specifically, one person (the speaker or "instructor") remotely guided another person (the listener or "walker") through a hall to a number of desks with distractors and target items with which different tasks had to be performed, such as assembling utensils for baking a cake. Both task and target items were only known to the instructor. The walker was eye-tracked and the instructor saw the output of the eye tracker's scene camera only (NoGaze), the video overlaid with the walker's gaze position (Gaze) or the video overlaid with the current gaze position to which we artificially added 20% random error (ManGaze).

While task performance did not vary with *GazeAvailability*, the amount of feedback given by the speaker did to some extent. We further found that listeners' gaze behaviour differed as a function of whether or not an utterance was taking place, probably reflecting language comprehension processes. Moreover, gaze patterns also changed with *GazeAvailability* to the speaker. In particular, we analysed scenes immediately *before* any utterance onset but also di-

rectly *after* utterance offset. We take the former to provide some indication for gaze being used as *signal* to which a speaker reacts, whereas the latter suggests that *GazeAvailability* may also have an indirect influence onto the speaker's utterances which, in turn, have an impact on listener gaze again.

Related work

Previous research has shown that listeners follow speakers' verbal references (as well as her gaze in face-to-face situations) to rapidly identify a referent (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Hanna & Tanenhaus, 2004; Keysar, Barr, Balin, & Brauner, 2000). The reaction of the speaker to such referential eye movements, however, was considered in few studies. Clark and Krych (2004), for instance, aimed to grasp this *reciprocal* nature of interaction in a study using a collaborative block building task and manipulating whether participants could see each other or each other's workspace. Their results suggested that the joint workspace was more important than, for instance, seeing each other's faces. Staudte, Koller, Garoufi, and Crocker (2012) conducted a study in which users were guided by a natural language generation (NLG) system through a virtual world to find a trophy. The system either gave feedback to the users' eye movements, or not. This controlled setting allowed the observation of dynamic and interactive (gaze) behaviour while maintaining control on one interlocutor (the NLG system). The results of this study suggest that it can be beneficial for task performance when listener gaze is exploited by the speaker to give feedback. It remains unclear, however, whether (human) speakers indeed provide such feedback and how the availability of listener gaze *recursively* affects the spoken instructions and, possibly, the gaze behaviour itself.

Experiment

We designed an experiment that combines a dynamic, interactive setting with the possibility to conduct exact and detailed analyses, in particular on eye movement behaviour, in order to assess the mutual influence of listener gaze and speech. Naïve participants either became an instructor (speaker) or a walker (listener). The speaker instructed the listener to perform a series of tasks. These tasks consisted of a navigational part, i.e. finding the next out of nine tables in a hall, which we call the *macro* task, and some object assembly at each table, the so-called *micro* tasks. Each pair of participants experienced all three *GazeAvailability* manipulations.

The listener wore a head-mounted eye tracker through which the speaker could see the scene from the listener's perspective (NoGaze) and additionally the listener's exact gaze cursor (Gaze) or a manipulated gaze point (ManGaze). The purpose of this manipulation *GazeAvailability* was to reveal whether the availability of listener gaze to the speaker affected a) the produced utterances and b) the listener's gaze. The purpose of including ManGaze was to investigate whether slightly perturbed gaze would be considered either un-informative or even disturbing (more like NoGaze), or whether the speaker would be robust towards slight imprecisions of the gaze cursor and treat it more like Gaze.

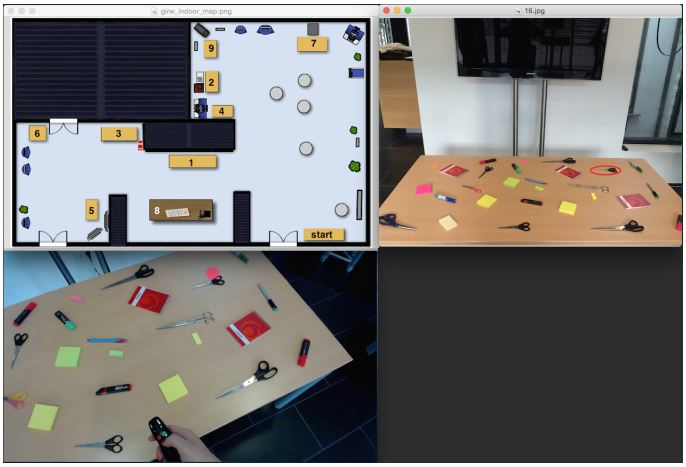

Figure 1: A screenshot of instructor's display.

Firstly, the experiment was aimed to reveal whether the availability of listener gaze position to the speaker would affect the production of verbal feedback. And secondly, if gaze was used as a *signal*, which listeners control and use deliberately, then the option to do so (and thereby evoke speaker reactions) would ubiquitously change listener gaze. If gaze was more generally a *symptom* of other processes and deliberate control was (too) difficult, listener gaze would rather change with tasks or events than with *GazeAvailability*. Finally, if gaze was used as a *signal*, variations of listener gaze behaviour should mainly occur prior to an utterance. If gaze was a reaction to changes in the utterances (i.e. a *symptom*), gaze behaviour should rather change after an utterance.

Method

The instructor received a plan of the route and a picture of the table top in which the next target object for the *micro* task was highlighted, see Figure 1. To make the task sufficiently complex and elicit precise references to target objects, at least two distractors for each target were also on the table.

The experiment consisted of nine micro tasks, each of which was dedicated to some everyday life activity such as office work or cooking. Office tasks included writing a letter using envelopes, pens, blocks and glue; kitchen tasks making a cake using milk, sprinkles, mixing spoons and an eggs box. In total, 234 every day objects were used, 36 of which were target objects.

Participants Twelve pairs of participants (16 females) took part in this study. Average age was 26.6 and all but one were in the age range 18-40. All participants were German native speakers and received a payment of 10 €. A session lasted between 30 and 45 minutes.

Procedure Participants were first asked about their preference for role assignment and assigned to the walker/instructor role accordingly. Two experimenters instructed both participants separately from each other. Specifically, the instructor was shown the route and tables but was not told how to refer to target objects. Then, she was led to a remote room from where she guided the walker. During the experiment the instructor saw a picture of the current target object, a map of the hall, and the scene view of the walker. Neither walker nor instructor were informed about our manipulation.

Apparatus

We used a Pupil Pro monocular head-mounted eye tracker for gaze data collection (Kassner, Patera, & Bulling, 2014). The tracker is equipped with a high-resolution scene (resolution of 1280 x 720 pixels) and eye camera (640 x 360 pixels). We extended the Pupil software with additional functionality needed for our study, namely to hide and display a manipulated gaze cursor to the instructor.

Two notebooks were used: one for the walker and one for the instructor. The instructor notebook was connected to two displays, one for the instructor and one for the experimenter. The experimenter sitting next to the instructor used a control panel to send commands to the eye tracking software to switch between conditions. The eye tracker was connected to the walker notebook on which we recorded the incoming sound, i.e. the instructions the listener heard. Both audio and video signals were streamed using Skype. In addition, the walker was equipped with a presenter to signal success (finding a target object) by pressing a green button which was used later on to segment the micro tasks. The communication of the different software components was implemented using a custom client-server software but all recordings were carried out on the walker machine.

Analysis

Linguistic analysis To prepare the recorded data for further processing, we applied a standard linguistic preprocessing pipeline. We first transcribed the audio signal, which was a manual step as the discourse collected in our study was very specific and contained also ungrammatical utterances. We then aligned the text to the audio signal by applying the forced alignment technique (Kisler, Schiel, & Sloetjes, 2012). We performed lemmatization and part-of-speech (POS) tagging followed by linguistic annotation using shallow syntactic analysis. These annotations are automatically carried out using TreeTagger (Schmid, 1995).

Two types of feedback instances, positive and negative, were recognized by searching for word occurrences that express feedback, e.g. "ja, genau" (*yeah right*) is positive while "nein, falsch" (*no, wrong*) is negative. However, in some cases those words did not express feedback but had a different grammatical function. Therefore a manual post correction was carried out to filter out detected instances and also to add few other words that are not typical for feedback but had this function in a particular context.

Lastly, we assessed the proportion of positive and negative feedback instances per condition. We used linear mixed-effects models using the lme4 package in R (Baayen, Davidson, & Bates, 2008) and model selection in order to determine the influence of *GazeAvailability*.

Eye movement analysis We first detected fixations using a standard dispersion-based fixation detection algorithm as in (Salvucci & Goldberg, 2000) that declares a sequence of gaze points to be a fixation if the maximum distance from their joint center is less than 5% of the scene camera width and the sequence has a minimum duration of 66 msec. Eye movements between two fixations were considered saccades without further processing. Blinks were not included as video-based eye trackers, such as Pupil, do not record them by default. We then used a sliding window approach with a window size of 500 msec and step size of 250 msec to extract eye movement features, resulting in a dataset consisting of 18841 time windows.

For each window, we extracted a subset of 45 features of those previously proposed for eye-based recognition of visual memory recall processes (Bulling & Roggen, 2011) and cognitive load (Tessendorf et al., 2011). We added 21 additional features relating current gaze behaviour to the overall gaze behaviour of the current person in the current experiment, e.g. the ratio of the small saccade rate in the whole experiment to the small saccade rate in this time window. All features are shown in Table 1. For feature selection we used the minimal-redundancy-maximal-relevance criterion (mRMR) which aims to maximise the feature's relevance in terms of mutual information between target variable and features while discarding redundant features (Peng, 2007). For our analyses we relied on data driven method and used the consistently top-ranked features for target variables such as *GazeAvailability*, *Pair* or *FeedbackPresence* and fitted linear mixed-effects models to the top-ranked feature according to mRMR (saccade rate). Similar results can also be achieved based on further top-ranked features such as the ratio of small to large saccades (where a saccade is considered small if its amplitude is less than twice the maximum radius of a fixation).

Results

We first evaluated the time needed to solve a task (all tasks were solved) in each condition to reveal whether gaze was used to complete a task more efficiently. It took participants 64.16 seconds on average to solve a task in the GAZE condition, 62.96 seconds in the MANGAZE condition, and 64.46 seconds in the NOGAZE condition. Differences were not significant. Since the average interaction time was generally very low, a floor effect has possibly prevented a distinction of the conditions.

Fixation	rate, mean, max, variance of durations mean, variance of variance within one fix.
Saccades	rate, ratio of (small/large/right/left) sacc. mean, max, variance of amplitudes
Combined	ratio saccades / fixations
Wordbooks	number of non-zero entries maximum and minimum entries as well as their difference for n-grams with $n \leq 4$
Ratios	all fixation, saccade and combined features in ratio to the value over the whole trial for a particular pair and condition.

Table 1: Features extracted from human visual behaviour inspired by Bulling et al. (2011).

Linguistic analysis

Next we examined the intuition that the length of utterances would be shorter in the GAZE condition and longer in the MANGAZE condition. However, no significant difference was found.

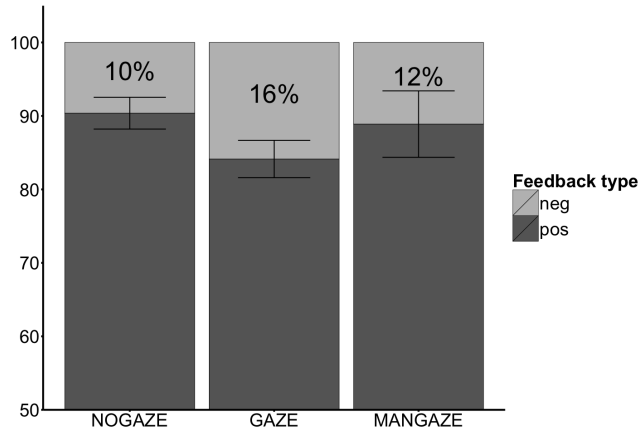

Figure 2: The proportion of positive and negative feedback instances in the different conditions. The model fitted to that data is the following `feedbackType ~ GazeAvailability + (1|Pair)`

We then investigated the proportion of positive and negative feedback. To test if the difference in the proportions is significant we constructed a generalised linear mixed-effects model (with a logit link function) fitted to *FeedbackType* with *GazeAvailability* as a fixed effect.

Figure 2 depicts a graph that shows the proportion of feedback in the different gaze conditions and gives the model specifications. The amount of data points (feedback instances per pair) does not license the inclusion of a random slope in the model so we include only the random intercept for *Pair*.

This model shows a difference between the GAZE and NOGAZE condition that approaches significance (Coeff. = 0.574; SE=0.314; Wald's Z=1.829; $p = .067$). This marginally significant difference indicates that speakers make use of the exact gaze positions of the listeners and that they utter more negative feedback to signal misunderstandings. MANGAZE is treated somewhat in-between GAZE and NOGAZE.

Moreover, a negative feedback instance is usually followed firstly by a repair, i.e. an additional description that either provides a complementary information that was not mentioned in the instruction before or an alternative description that describes a distractor which is usually underspecified. Secondly, a positive feedback instance often follows to confirm the successful resolution of the repair. Example (1) illustrates that repeated pattern.

(1) ne das andere ... Genau (*no the other one ... exactly*)

We further explored if these repairs differed with the availability of gaze: We measured the length (in words) of the repairs and compared them across all conditions but found no significant differences.

Visual behaviour analysis

To asses the role of listener gaze in this scenario, we examined the interplay of utterances, listener gaze and the *GazeAvailability* manipulation.

First, we fitted a linear mixed-effects model with a random intercept and random slope for *pair* to the data set consisting of all (sliding) time windows (18841 in total). We found a significant main effect of *UtterancePresence* through model selection ($\chi^2(1) = 9.54$, $p = .002$). *GazeAvailability*, in contrast, had no effect on model fit. We then considered feedback expressions which are a specific form of utterance and commonly occur in situated and spoken interaction: Such phrases typically form a direct and closely time-locked response to changes in the situation or, more crucially, the listener's behaviour. Similarly to the analysis of utterances in general, we fitted a linear mixed-effects model, this time with *FeedbackPresence* as a factor. We observed a main effect ($\chi^2(1) = 80.63$, $p < .001$) and an interaction with *GazeAvailability* ($\chi^2(2) = 9.38$, $p = .009$). The interaction suggests that the manipulation of gaze availability has some effect on how listeners move their eyes during verbal feedback compared to before or after. This observation also seems to be in line with the results of the linguistic analysis according to which the proportion of positive and negative feedback instances vary in the different levels of *GazeAvailability* to the speaker.

Taken together, the results from gaze behaviour in *UtterancePresence* and *FeedbackPresence* indicate that gaze patterns differ with speech happening or not, i.e. when the listener is processing speech compared to when she is not currently listening to an utterance, and that this is relatively independent of *GazeAvailability*. In light of the symptom-signal-distinction, this suggests that language comprehension processes drive the ocular system (*symptom*) but that deliberate control of gaze, e.g. using it as pointer in the GAZE but not the NOGAZE condition (*signal*), hardly affects overall gaze

patterns.

Furthermore, we attempted to break up the reciprocal nature of the interaction between listener gaze and speech by considering the temporal order of gaze events and speech events. Examining how gaze affect utterances and then, in turn, how the utterances affect eye movements helps us to shed light onto the dual role of listener gaze: On the one hand, it can be seen as a sign that helps the walker to communicate with the instructor (as the instructor can observe the walker's behaviour but cannot hear the walker). In this case, gaze patterns may differ between the GAZE and NOGAZE conditions *before* an utterance, since in the former condition gaze may be more frequently used as a *signal* to which the speaker reacts. On the other hand, gaze may be mostly a *symptom* that reflects language processing and which therefore may also reflect when the speaker adapts to seeing listener gaze (GAZE condition) and produces utterances accordingly. In that case, gaze patterns are likely to differ with *GazeAvailability* immediately *after* utterance offset.

Thus, analogously to the analyses above, we fitted linear mixed-effects models on a subset of the data, namely the time windows immediately *before* the onset and *after* the offset of an utterance. Both subsets consist of 954 instances and we found that the factor *GazeAvailability* significantly contributes to a better model fit, not only *before* an instruction ($\chi^2(2) = 9.77$, $p = .008$) but also *after* it ($\chi^2(2) = 10.89$, $p = .004$). The same analysis was carried out for the time windows *before* and *after* positive and, additionally, *before* and *after* negative feedback occurrences. However, no effect of *GazeAvailability* was observed (which may also be due to the lower number of samples).

To conclude, we observed no significant difference in gaze behaviour along with the *GazeAvailability* manipulation, but gaze patterns were distinct from each other in the presence and absence of utterances in general and feedback in particular. The analyses taking temporal aspects of the gaze and speech events into consideration showed that listener gaze significantly differs *before* and *after* instructions. This evidence supports the view that listener gaze can not only be seen as a *symptom* of language comprehension but also a non verbal *signal* to the speaker. The latter role is comparable to the role of verbal deictic expression like "Do you mean this one there?" which may have been used in a bidirectional verbal dialogue.

Discussion

In this exploratory study, we observed that the manipulation of availability of listener gaze position to the speaker had a main effect on listener gaze *before* and *after* an utterance, but not while an instruction was being spoken. *GazeAvailability* further affected the type and amount of feedback given by speakers. In particular, GAZE differed significantly from NOGAZE with MANGAZE being in-between those two condition with respect to the amount of negative feedback uttered by the speaker. This suggests that manipulated gaze was used somewhat less than natural gaze but was not ignored either.

Based on the combination of gaze effects *before* and *after* an utterance and the lack of such an effect on eye movements *during* an utterance, we further assume that listener gaze can be seen as both a *signal* from listeners for conveying some sort of information to the speaker and as *symptom* that reflects language comprehension processes.

The tendency of speakers to produce more negative feedback with gaze availability also supports the role of listener gaze as a *signal* to which instructors actively react. These feedback instances have the potential to quickly eliminate wrong beliefs by the listener about intended referents. We did not find an improvement of task performance in terms of time needed for completion in the GAZE condition but we believe that this could be due to a ceiling effect.

Similarly, we did not find a significant effect of *GazeAvailability* on other coarse-grained measures of the spoken material such as utterance length (in words). However, many words do not necessarily carry more information. Importantly, the salience threshold for the speech segmentation is also a crucial parameter and can vary depending on the domain, task and setting, e.g. whether it is an uni- or bidirectional, free or goal-oriented conversation. In addition, the word level may be too coarse to reveal qualitative differences in utterances as a function of listener gaze as mentioned in Section *Linguistic analysis*. Hence it may be worth examining whether the instructions collected in the recorded interaction can be distinguished on a more fine-grained linguistic layer but this was beyond the scope of this paper.

Future Work

There are several caveats in this study which motivate future work. Firstly, the possibility for listeners to show their hands, make pointing gestures or hover over objects probably added noise to the role of listener gaze as a feedback modality. A follow-up study in a virtual environment will avoid this and hopefully increase clarity of the gaze-speech interaction pattern. Secondly, the experiment consisted of a micro and a macro scale task, the latter of which was originally intended to be more of a navigation task. The actual reduction in task complexity (and therefore for the neglect in the analyses) lies in the significant technical challenges to set up a stable WLAN connection throughout a large building to transfer high-resolution video data, audio, and gaze data in real time. Thirdly, we considered relatively coarse, quantitative measures for utterances so far, mostly due to the lack of manpower in annotating the data. Further analyses with respect to such richer annotations as well as other eye movement analyses (such as using smooth pursuit) are planned. Lastly, we plan to classify scenes as containing confusion or misunderstandings which would be of particular interest, for instance, for a machine learning approach in order to detect confusion from eye movements. Measures of cognitive load, for instance, may be extracted from the available data to label scene segments accordingly.

Conclusion

We reported on an indoor guidance study to explore the interplay between spontaneous spoken instructions and listeners' eye-movement behaviour. We presented a study design and experimental setup to collect a multi-modal dataset of scene view videos from the listener's perspective, their gaze data, and instructors' verbal instructions. We found that listener gaze itself as well as the speaker's utterances were affected by *GazeAvailability*. The specific pattern of effects suggests that listener gaze is not only a processing *symptom* that is affected indirectly by the variation of *GazeAvailability*, but also as a *signal* being used deliberately as a pointing gesture.

To conclude, we have presented an exploratory study which aimed to shed some light on the role of listener gaze (position) in an interactive, indoor guidance task. The study combines an interactive and very dynamic setting with fine-grained data collection and analyses. The presented results can be seen as a first step towards understanding the reciprocal nature of gaze behaviour and speech in human interaction which may, for instance, help artificial interaction partners to exploit human gaze in making communication more efficient and less error-prone.

Acknowledgements

We would like to thank Torsten Jachmann and Laura Faust for carrying out the annotations and Ross Macdonald for valuable discussions. Many thanks also to very helpful and encouraging anonymous reviewers. This work was funded by the Cluster of Excellence on "Multimodal Computing and Interaction" of the German Excellence Initiative and by the German Ministry of Education and Research (project MOBIDA—AD; grant number 01IS12050).

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008, November). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. Retrieved from http://dx.doi.org/10.1016/j.jml.2007.12.005 doi: 10.1016/j.jml.2007.12.005

Bulling, A., & Roggen, D. (2011). Recognition of Visual Memory Recall Processes Using Eye Movement Analysis. In *Proc. UbiComp* (p. 455-464).

Bulling, A., Ward, J. A., Gellersen, H., & Tröster, G. (2011). Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *33*(4), 741-753.

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

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84-107.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*(6), 409–436.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cognitive Science*, *28*(1), 105 - 115.

Kassner, M., Patera, W., & Bulling, A. (2014). Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Adj. Proc. UbiComp* (p. 1151-1160). Retrieved from http://pupil-labs.com/pupil/

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: the role of mutual knowledge in comprehension. *Psychological Science*, *11*, 32–38.

Kisler, T., Schiel, F., & Sloetjes, H. (2012). Signal processing via web services: the use case webmaus. In *Proceedings digital humanities 2012, hamburg, germany* (p. 30-34). Hamburg.

Peng, H. (2007). *mRMR Feature Selection Toolbox for MATLAB, http://penglab.janelia.org/proj/mrmr/*. Retrieved from http://penglab.janelia.org/proj/mRMR/

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proc. ETRA* (pp. 71–78).

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *In proceedings of the acl sigdat-workshop* (pp. 47–50).

Staudte, M., Koller, A., Garoufi, K., & Crocker, M. W. (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.* Sapporo, Japan.

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

Tessendorf, B., Bulling, A., Roggen, D., Stiefmeier, T., Feilner, M., Derleth, P., & Tröster, G. (2011). Recognition of hearing needs from body and eye movements to improve hearing instruments. In *Proc. Pervasive* (pp. 314–331).

Yantis, S., & Jonides, J. (1990). Abrupt visual onsets and selective attention: Voluntary versus automatic allocation. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 121–134.