

An Eye for an Eye, and for Other Modalities

Max M. Louwerse (mlouwerse@memphis.edu)
Arthur C. Graesser (a.graesser@mail.psyc.memphis.edu)
Danielle S. McNamara (d.mcnamara@mail.psyc.memphis.edu)
Gwyneth Lewis (glewis@mail.psyc.memphis.edu)
Megan Zirnstein (mzirnstn@memphis.edu)

Department of Psychology / Institute for Intelligent Systems
Memphis, TN 38152 USA

Abstract

Many eye tracking studies in psycholinguistics have investigated reading processes at sentential and intersentential levels. Relatively few have investigated what participants' eyes do during common psycholinguistic scenarios like face-to-face interactions. We will give an overview of four eye tracking studies that investigated aspects of multimodal communication. The first pair of studies investigated the role of gestures in communication, one looking at human-human scenarios, the other at human-computer scenarios. The second pair looked at face-to-face communication, again one looking at human-human scenarios, the other at human-computer scenarios. This overview aims to show the importance of eye tracking technology in studies investigating multimodal communication.

Introduction

Imagine asking somebody for directions to a particular location. While your dialogue partner is busy talking, you are equally busy. While your partner is talking, you are working hard to not only understand the linguistic input, but to also make eye contact, maintain eye contact, and pay attention to any deictic gestures your dialogue partner makes in order to identify the directions you need to take.

Eye gaze has proven to be extremely important in face-to-face interactions. Within the scope of this paper, a brief overview of studies that investigated eye gaze in communication needs to suffice. There is evidence that blinks are synchronized to the word or even the syllable level (Argyle & Cook, 1976) and are used to emphasize speech, accent a word, or mark a pause (Ekman, 1979). Eyes also signal turn boundaries in conversations (Novick, Hansen, Ward, 1996). The communicative value of eye gaze is such that it can replace explicit verbal cues (Doherty-Sneddon, et al., 1997).

Psycholinguistic research on multimodal communication has, however, been scarce. For example, in the Hyönä, Radach, and Deubel (2003) overview of cognitive and applied eye movement research only one of the 33 included papers was devoted to eye movement in face-to-face interaction. The vast majority of the papers, however, considered eye tracking in reading only natural texts, graphics, or advertisements. This is a pity for a number of reasons. First, eye tracking technology can provide us with important insights into the most natural forms of communication, like face-to-face conversation. Secondly,

these insights can be extended for many human-computer applications in which face-to-face scenarios are modeled using animated conversational agents. Without these insights it is unlikely that animated conversational agents will ever be perceived as natural. Thirdly, eye tracking is probably the only methodology successful in investigating face-to-face conversations. Reading research can also acquire much insight from self-paced, moving-window reading time (Graesser, Hoffman, & Clark, 1980), priming experiments (T. McNamara, 2005), or read-aloud and think-aloud protocols (Bereiter, 1985), perhaps though with less precision and in less natural contexts than eye tracking. Measuring what participants pay attention to in face-to-face scenarios, however, does not have a good alternative methodology than eye tracking technology.

This paper gives an overview of four studies that have been conducted in our eye tracking labs. The first pair of studies investigated how deictic gestures and speech interact when a dialogue partner points at a target, linguistically and/or non-linguistically. The first study looked at human-human scenarios, the second at human-computer scenarios. The second pair of experiments investigated what dialogue participants pay attention to in face-to-face interactions. Do they look at the dialogue partner? Do they look at objects that form part of the communication, like maps, diagrams, and pictures? As in the first pair of studies, the first study looked at human-human scenarios, the second at human-computer scenarios.

1. Paying Attention to Gestures

In two eye tracking studies, Louwerse and Bangerter (2005; under review) investigated the interaction between pointing gestures and spatial descriptions like "here", "there", "right at the top". More specifically, they investigated whether these deictic gestures are substitutable for these deictic expressions and whether these pointing devices establish joint attention between speaker and hearer. Both studies aimed to determine whether pointing helps the hearer in the communicative process. The hypothesis under investigation was whether deictic gestures help hearers identify the target indirectly, by guiding their gaze to its region. By this hypothesis, pointing helps establish a joint focus of attention between speaker and hearer (joint-attention hypothesis). Second, the Louwerse and Bangerter studies aimed at determining whether deictic gestures are substitutable for

certain linguistic spatial expressions (substitution hypothesis). The first eye tracking study tested these hypotheses in a human-human scenario, the second in a human-computer scenario.

1. Human-Human Interaction

Louwerse and Bangerter (2005) investigated the effects of referring expressions and pointing gestures on the addressee’s attention. Participants saw short movies (5 seconds each), each movie consisting of 12 smiley faces differing in props and emotion and, depending on the condition, a human pointer, pointing out and/or describing the target. Every participant cycled through each of the six conditions five times in random order. Four of these conditions are presented in Table 1.

Table 1: Overview of pointing x description conditions

	location description	no location description
pointing	Pointing + John is on the top left with a hat and bow tie	Pointing + John has a hat and bow tie
no pointing	John is on the top left with a hat and bow tie	John has a hat and bow tie

Two additional conditions were created by manipulating the time and order of linguistic expressions and pointing: in a fifth condition, pointing preceded the linguistic expressions (feature description only), but with an inserted pause of two seconds. In the final condition the feature description followed the pointing after a two-second pause. Participants’ eye movements were tracked using a Model 501 Applied Science Laboratory eye tracker with temporal resolution of 60 Hz and a spatial resolution of .50 degree angle horizontally and a .40 degree angle vertically.

Participants were asked to watch each clip and click the mouse button as soon as they had identified the target face. The eye tracking and accuracy findings supported the view that deictic gestures can substitute for language functions. That is, when a feature description was accompanied by either a deictic gesture or a deictic expression, accuracy in target identification increased. However, when both the deictic gesture and the deictic expression were present, no additional gains are found in accuracy. This pattern was also found in the number of regressive eye movements. Participants spent more time on the correct target when pointing was present or when the location description was present, but not when they were combined. This provided support for what we called the substitution hypothesis for deictic gestures.

The results also provided evidence for the joint-attention hypothesis, which states that gestures support communicative joint activities. Pointing was shown by the eye tracking data to help establish a joint focus of attention between speaker and hearer. If the joint focus of attention was identified after the target identification, it results in confusion, as indicated by more regressive eye movements, higher fixation times to identify the target, and more non-

targets being considered before the correct target is identified.

1.2 Human-Computer Interaction

The second experiment (Louwerse & Bangerter, under review) again tested the substitution and joint-attention hypotheses except for two differences with Experiment 1. First, an artificial environment was used wherein pointing gestures and linguistic expressions were generated artificially. Second, all conditions used in Experiment 1 were maintained, except the gesture-feature delay condition was replaced by a linguistic deixis condition to compare the role of different linguistic expressions with deictic gestures.

The same experimental design was used as in Experiment 1, with each participant being exposed to all six conditions. Five of these conditions were identical to the four presented in Table 1 and the delay condition in which the gesture followed the speech after a two second delay. The condition different from Experiment 1 consisted of the spatial description being replaced by a deictic expression (John is over there with a hat and bow tie) to determine the effect of general deictic expressions (over there) with specific location expressions (John is on the top left with a hat and bow tie) used in Experiment 1.

Participants’ eye movements were tracked using an SMI iView X High-Speed eye tracker with a temporal resolution of 240Hz. The horizontal viewing angle was + 30° and vertical view angle was 30° up and 45° down. The procedure for Experiment 2 was identical to that of Experiment 1.

Results showed patterns identical to those found in Experiment 1. The artificial synthesized speech and ClipArt pointing thereby resulted in equal, if not higher, performance accuracy as the human speech and pointing. As in the previous experiment, results showed that comprehension was facilitated when general pointing preceded specific pointing. However, comprehension was hindered when pointing was followed by general information. Both the human-human and the human-computer eye tracking studies thus provided evidence for the joint-attention and the substitution hypotheses. Eye tracking provided an invaluable methodology to determine online processes in target identification.

2. Paying Attention to the Dialogue Partner

The eye tracking studies reported so far looked at gestures in multimodal communication. Carefully manipulated stimuli were used, whereby the participant could not see more than a hand pointing at one of the presented targets.. Eye tracking can also, of course, be used in more natural scenarios of multimodal communication. We briefly describe two eye tracking experiments here that have done exactly that. The first experiment investigates the role of eye gaze in multimodal communication between humans, the second between humans and animated conversational agents.

2.1 Human-Human Interaction

In an ongoing project on multimodal communication in humans and agents (Louwerse, Bard, Steedman, Graesser & Hu, 2004), we are investigating the interaction between linguistic modalities, such as prosody and dialogue structure, and non-linguistic modalities, such as eye gaze and facial expressions. The project aims to determine how these modalities are aligned, whether, and if so when, the interlocutor observes these modalities, and whether the correct use of these channels actually aids the interlocutor's comprehension. Answers to these questions should provide a better understanding of the use of communicative resources in discourse and can subsequently aid the development of more effective animated conversational agents.

With so many variables in multimodal communication, it is desirable to control for genre, topic, and goals of unscripted dialogs. Therefore, we used the Map Task scenario (Anderson, et al., 1991) because it provides a restricted-domain, route-communication task that makes clear to experimenters exactly what each participant knows at any given time. The Instruction Giver (Giver) coaches the Instruction Follower (Follower) through a route on the map. By way of instruction, participants are told that they and their interlocutors have maps of the same location but drawn by different explorers and so are potentially different in detail. They are not told where or how the maps differ. The maps are based on fictional locations.

Louwerse, Jeuniaux, Hoque, Jie, and Lewis (2006) reported the first results of these multimodal communication experiments. Cognitive states, eye gaze, and pauses were compared within both Giver and Follower, except for eye gaze, which was only recorded for the Giver. Eye gaze of the Follower was not recorded for the simple reason that only one eye tracker was available. Recording of the Giver's eye gaze was considered most important because of the development of the animated conversational agent. Cognitive states were determined by facial expressions. These were inspired by the Facial Action Coding System of Ekman and Friesen (1978) and included Distraction, Uncertainty, Confusion, Frustration, Confidence, Engagement, Encouragement, Interest, and Boredom. Facial expressions were coded by agreed-upon specifications and time stamped for particular occurrence.

Results showed that the Givers' eye gaze on the Follower correlated significantly with the cognitive states of Engagement, Uncertainty, and Boredom measured through facial expressions. In fact, Engagement and Uncertainty also correlated with the fixations on the map. Givers heavily involved in the task seemed to pay more attention to both their dialogue partner and the map in front of them, either because they were absorbed in the task or because they were uncertain about an aspect of that task. The same patterns were also found for the Follower. Though the Followers' eye gaze was not recorded, the coding for the Follower moving his or her eyes away from the map gave an adequate approximation of fixation on the Giver. Again, Uncertainty

and Engagement were the cognitive states during which this happened most frequently.

Eye blinks co-occurred with many of the cognitive states. It is noteworthy that, in addition to cognitive states like Engagement and Uncertainty, blinks correlated with the cognitive state of Confusion. On the contrary, when Givers felt confident, they paused less and looked less at the Follower, as suggested by significant negative correlations.

2.2 Human-Computer Interaction

How does the role of eye gaze in human-human communication carry over to human-computer communication, for instances in cases where a computer users communicates with an animated conversational agent? It is a human tendency to confuse what is real with what is perceived to be real, and we, therefore, automatically use social rules to guide our actions. Consequently, people interpret computers as social partners (Reeves & Nass, 1996). Several studies have shown that animated conversational agents are interpreted in this way in particular, so much so that participants activate stereotypes when interacting with agents (Louwerse, Graesser, Lu & Mitchell, 2005). In two eye tracking studies Louwerse, Graesser, McNamara, Bell and Lu (under review) monitored the eye gaze of participants.

In the first experiment, participants interacted with the intelligent tutoring system AutoTutor. AutoTutor is an intelligent conversational tutoring system that assists students in actively constructing knowledge by holding a conversation using natural language (Graesser et al., 2004; Graesser, VanLehn et al., 2001). AutoTutor poses questions or problems that require approximately a paragraph of information from a student as an answer. The interaction between the system and the user enables AutoTutor to engage the learner in a dialogue that assists in the evolution of an improved answer and draws out more of the learner's knowledge. Results showed that the AutoTutor agent served attracted subjects' attention and that this effect did not wear off over time.

To eliminate the explanation that the findings of Experiment 1 may be attributed to the fact that there is one colorful animation on the screen (the AutoTutor agent) and that this may have guided a participant's attention, a second eye tracking experiment with multiple agents was conducted. Participants were presented with the introduction of the Interactive Strategy Training for Active Reading and Thinking (iSTART) (McNamara et al., 2004). iSTART provides training to help students more effectively self-explain difficult texts while reading. The system is based on experimental evidence that self-explanation coupled with reading strategy training (Self-Explanation Reading Training) improves comprehension and course grades (McNamara, 2004). iSTART delivers reading strategy training using an interactive and adaptive format. In the introduction phase of the system, three ECA's interact with each other and with the student to increase active processing and participation. The results showed that participants look

at agents when they speak and look away when they do not, despite the fact that nothing requires them to look at the agent to comprehend the delivered information.

Both eye tracking experiments showed that the face of a speaker – human or agent – attracts attention.

Conclusion

The four studies reported here illustrate the role of eye tracking technology in multimodal communication. The first pair of eye tracking studies investigated eye gaze and the role of gesture in communication, one focusing on human-human, the other on human-computer interaction. Results demonstrated the importance of gesture, as they can substitute for linguistic expressions. Moreover, no differences were found between human generated and computer generated speech and gestures. The second pair of eye tracking studies investigated eye gaze in multimodal communication, the first study focused on human-human, the second on human-computer interaction. Results demonstrate the complex interplay of modalities, including eye gaze, in the communicative process. Moreover, eye gaze on animated conversational agents was very similar to patterns found for human-human interaction.

It is hard to imagine methodologies other than eye tracking that would allow us to obtain comparable data and insights into multimodal communication between humans and between humans and agents. Even though eye tracking technology has been very limited in this area, it is worth keeping an eye on communicative modalities.

Acknowledgments

This research was supported by grants from National Science Foundation (IIS-0416128, ITR 0325428, REC 0106965, REC 0126265, REC-0089271) and IES (R305G040046). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF or IES.

References

- Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. & Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, 34, 351-366.
- Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.
- Bereiter, C. (1985). Use of thinking aloud in identification and teaching of reading comprehension strategies. *Cognition and Instruction*, 2, pp. 131-156.
- Doherty-Sneddon, G., Anderson, A. H., O' Malley, C., Langton, S., Garrod, S., & Bruce, V. (1997). Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance. *Journal of Experimental Psychology: Applied*, 3, 105-125.
- Ekman, P. (1979). About brows: Emotional and conversational signals. In M. von Cranach, K. Froppa, W. Lepenies, & D. Ploog (Eds.), *Human ethology: Claims and limits of a new discipline: Contributions to the colloquium* (pp. 169-248). Cambridge: Cambridge University Press.
- Ekman, P. & Friesen, W.V. (1978). *Facial action coding system*. Palo Alto, CA: Consulting Psychologist Press.
- Graesser, A. C, Hoffman, N. L., & Clark, L. F. (1980). Structural components of reading time. *Journal of Verbal Learning and Verbal Behavior*, 19, 135-151.
- Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: a Tutor with Dialogue in Natural Language. *Behavior Research Methods, Instruments, and Computers*, 36, 180-193.
- Graesser, A. C., VanLehn, K., Rosé, C., Jordan, P., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine*, 22, 39-51.
- Hyönä, J., Radach, R., & Deubel, H. (2003). *The mind's eye: Cognitive and applied aspects of eye movement research*. Amsterdam, The Netherlands: North-Holland.
- Louwerse, M. M., Bangerter, A. (2005). Focusing attention with deictic gestures and linguistic expressions. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*.
- Louwerse, M. M., Bangerter, A. (under review). *Getting the point. The interaction between deictic gestures and expressions*.
- Louwerse, M. M., Bard, E. G., Steedman, M., Hu, X., Graesser, A.C. (2004). *Tracking multimodal communication in humans and agents*. Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN.
- Louwerse, M. M., Graesser, A.C., Lu, S., & Mitchell, H. H. (2005). Social cues in animated conversational agents. *Applied Cognitive Psychology*, 19, 1-12.
- Louwerse, M. M., Graesser, A. C., McNamara, D. S., Bell, C., Lu, S. (under review). *Paying attention to embodied conversational agents*.
- Louwerse, M. M., Jeuniaux, P., Hoque, M. E., Wu, J., Lewis, G. (2006). Multimodal communication in computer-mediated map task scenarios. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.
- McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, 38, 1-30.
- McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. New York, Psychology Press.
- McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, and Computers*, 36, 222-233.
- Novick, D. G., Hansen, D. G., & Ward, K. (1996). Coordinating turn-taking with gaze. *Proceedings of the International Conference on Spoken Language Processing* (pp. 1888-1891). Philadelphia.
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge, MA: Cambridge University Press.