Gaze matching of referring expressions in collaborative problem solving

Naoko KURIYAMA[1], Asuka TERAI[1], Masaaki YASUHARA[1],Takenobu TOKUNAGA[1], Kimihiko YAMAGISHI[1] and Takashi KUSUMI[2]
Tokyo Institute of Technology[1], Japan, Kyoto University[2], Japan
kuriyama@hum.titech.ac.jp

Abstract. Richardson and Dale (2005) showed that eye gaze matching between speakers and listeners contributed to language comprehension. While their study used a static image as a visual stimulus, and the speech and eye gaze of speakers and that of listeners were recorded serially, we recorded speech in synchronisation with eye gaze of both participants simultaneously in a collaborative problem solving setting. The analysis of the collected data revealed that the eye gaze matching rate is higher in successful pairs than in unsuccessful pairs, and the peak of the matching rate comes at different position from the onset of referring expressions depending on surface form of the expressions.

Introduction

Identifying objects in conversation is a fundamental human capability necessary to achieve efficient and successful collaboration on any real world task. To denote an intended object, linguistic expressions called *referring expressions* are used. They are realized in various forms in actual communication, such as pronouns, definite and indefinite noun phrases and demonstratives. The object which a referring expression denotes is called a *referent*. To identify the referent of a referring expression, various contextual information should be shared between speakers and listeners. Such information is called "common ground", which is constructed through "grounding" during the course of conversation (Clark and Schaefer (1989)). This common ground is crucial for resolving referring expressions and consequently for

successful achievement of collaboration tasks.

In recent eye movement research, Richardson and Dale (2005) investigated the relation between a speaker's and a listener's eye movement and the listener's comprehension. They asked a subject (a speaker) to tell about TV show characters by showing their pictures. The audio of her speech and eye movement were recorded. Then another subject was asked to listen to the speaker's audio recording while seeing the same pictures and then to answer a comprehension test. The listener's eye movement was also recorded while she listened to the speaker's speech. They conducted a cross-recurrent analysis of the speaker's and listener's eye movement and concluded that a coupling between their eye movement could be a good indicator of the success of their communication.

The present study investigates the relation between eye gaze matching of dialogue participants and the success of collaborative problem solving. Furthermore, we particularly focus on eye gaze matching triggered by referring expressions.

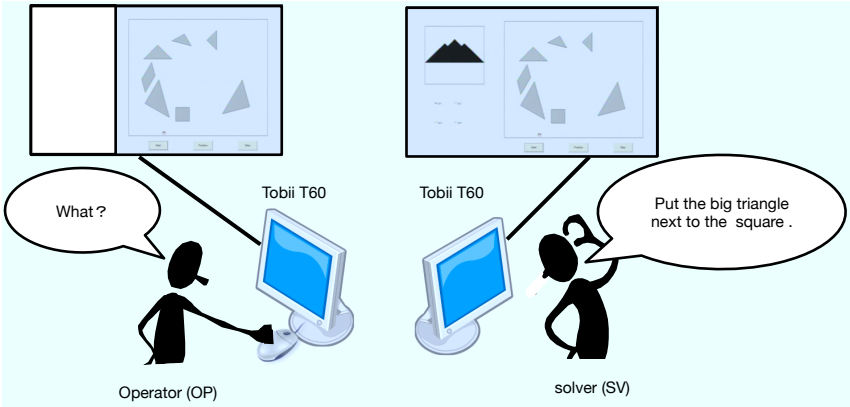


Figure 1. Experimental setting.

Method

An experiment of collaborative problem solving was conducted. The experimental setting is basically the same as Spanger et al. (2010) except that we also recorded the eye gaze of both participants in synchronization with participants' utterances and actions during problem solving.

Participants: We recruited 10 undergraduates and graduates to make 5 pairs of friends of the same gender. They are paid for participating in the experiment.

Apparatus: Participants worked on a Tangram simulator which shows a goal shape and a working area on the computer screen, and enables the participants to manipulate puzzle pieces by mouse operations. Two eye-trackers (Tobii T60) were used to record each participant's eye gaze.

Procedure: Each pair was instructed to collaboratively solve Tangram puzzles on the Tangram simulator. The goal of a Tangram puzzle is to construct a given goal shape by arranging all seven pieces in the working area as shown in Figure 1. Each pair is assigned a different role: a *solver* and a *operator* (Figure 1). The operator has a mouse for manipulating Tangram pieces, but does not see a goal shape on the screen. The solver sees a goal shape on the screen but does not have a mouse. This setting naturally leads to a situation where given a certain goal shape, the solver thinks of the necessary arrangement of the pieces and gives instructions to the operator how to move them, while the operator manipulates the pieces with the mouse according to the solver's instructions. They sat side by side with their own computer display showing the shared working area in real time. A room-divider screen was set between the solver and operator to prevent the operator from seeing the goal shape on the solver's screen, and to restrict their interaction to free speech only.

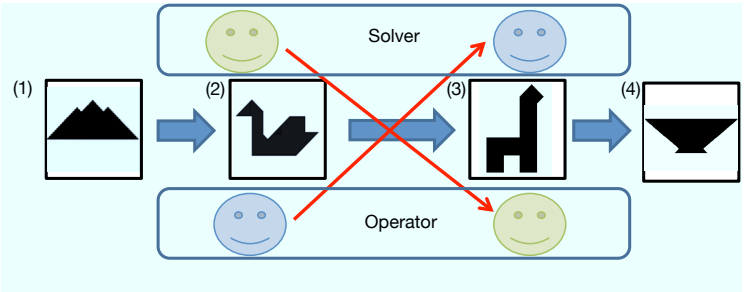


Figure 2. Goal shapes to solve.

Each pair was assigned 4 exercises (symmetric (1), (4) and asymmetric (2), (3)) as shown in Figure 2, and changed their roles after two exercises (1) and (2). Before starting the first exercise as the operator, each participant had a short training session in order to learn how to manipulate pieces with the mouse. The initial arrangement of the pieces was randomized every time. We set a time limit of 15 minutes for the completion of each goal shape.

A trial ends when the goal shape is complete or the time is up. Utterances by the participants are recorded separately in stereo through headset microphones in synchronization with the position of the pieces, the mouse operations and eye gaze of both participants. Piece positions and mouse actions were automatically recorded by the simulator at intervals of 1/65 second. A participant's gaze was captured by the Tobii T60 eye tracker at intervals of 1/60 second. The 9-point calibration of the eye tracker for both participants was conducted before starting the training session. The display size is 1,280 x 1,024 pixels and the distance between the display and participant's eye was maintained at about 45cm.

Result and Discussion

The data of two pairs (8 dialogues) whose eye gaze was successfully captured more than 70% of the total duration of a trial was used for analysis. The collected gaze data was smoothed by a low-pass filter of 6Hz. To see the extent of gaze matching, the cross-recurrence analysis was conducted (Richardson and Dale (2005)). The eye gaze of a pair was considered "matching" when the gaze of both participants stayed within the range of 100 pixels for more than 0.1 second.

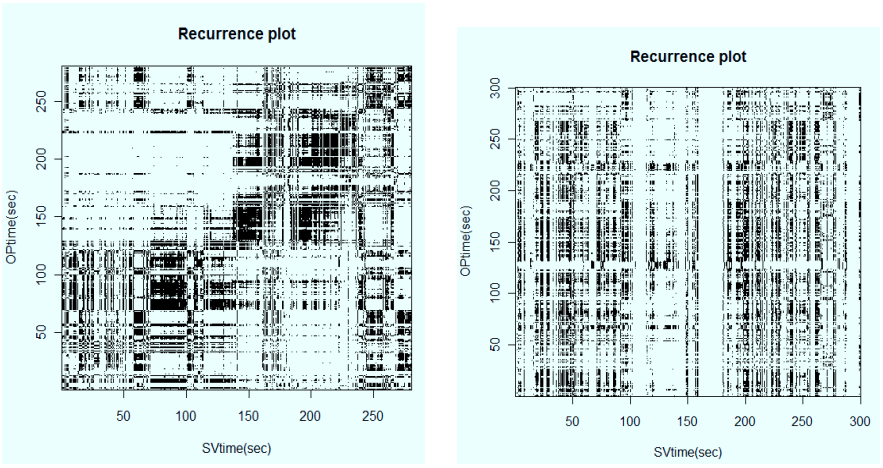
Pair	Goal shape	Success	Task comp. time [sec]	Gaze matching rate			
				total	early phase	middle phase	late phase
A	1	yes	886	0.19	0.11	0.26	0.21
A	3	yes	841	0.39	0.23	0.43	0.52
A	4	yes	697	0.40	0.37	0.42	0.42
B	4	yes	427	0.30	0.17	0.34	0.40
Average			713	0.32	0.22	0.36	0.39
A	2	no	–	0.20	0.18	0.24	0.18
B	1	no	–	0.31	0.28	0.30	0.35
B	2	no	–	0.30	0.33	0.26	0.30
B	3	no	–	0.32	0.20	0.19	0.38
Average			–	0.28	0.25	0.25	0.30

Table I. The eye gaze matching rate.

We investigated gaze matching between two participants of pairs with respect to the successful and unsuccessful pairs, the difference of goal shapes, and temporal phases in a trial. A trial period was divided into three phases of an equal time span: the early, the middle and the late phase. Table I shows the gaze matching rates of the combination of these factors. The gaze matching rate is calculated by the ratio of the sum of gaze matching periods in a trial to the total time of the trial. The rate tends to be higher in the pairs who successfully solved the puzzle within the time limit than in unsuccessful pairs except for the early phase. Since we do not have enough volume of data at the moment, statistical tests are difficult to perform.

The gaze data were examined by cross-recurrence analysis Zbilut et al. (1998) as well. When the gaze was on the diagonals of a plot (Figure 3), it indicated complete gaze correspondence between two participants, suggesting that they were sharing information well. Results indicated that the rate of gaze matching in the late phase of the successful pairs was higher than that of the unsuccessful pairs, which corroborated the previous study. This could be explained that both participants focus on unsolved parts of the goal shape near task completion, thus the region of their visual attention converges.

The cross-recurrence plot provides a good overview of gaze matching between participants from a macroscopic view. It is not, however, suitable for more precise analysis, since it might include the gaze matching regardless of the dialogue contents, i.e. it is difficult to discern the difference between meaningful and coincident gaze matching. Therefore, we conducted a microscopic analysis as well,



(a) successful case (561-841 sec) (b) unsuccessful case (601-900 sec)

Figure 3. Examples of cross-recurrence plot (successful vs. unsuccessful cases).

in particular, an analysis of gaze matching triggered by referring expressions. The research of the relations between eye gaze and human referring behavior has a long history (Ehrlich and Rayner (1983); van der Meulen et al. (2001); Metzing and Brennan (2003); Sturt (2003); Hanna and Tanenhaus (2004); Brown-Schmidt and Tanenhaus (2006); Hanna and Brennan (2007)), and referring to objects, particularly referring to puzzle pieces in our current study, is crucial to the successful completion of the task. Our experimental data includes 690 referring expressions denoting a puzzle piece, which were annotated by hand. Among these expressions, 517 were produced by the solvers and 173 by the operators. Each expression is also annotated with its referent and several attributes following the annotation schema by Tokunaga et al. (2010).

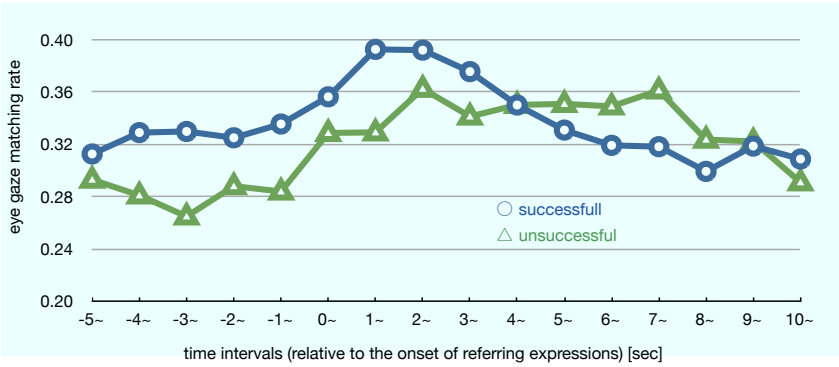


Figure 4. The eye gaze matching rate (successful vs. unsuccessful pairs).

We investigated eye gaze matching of the participants around the onset of uttering referring expressions. More specifically, we calculated the eye gaze matching rate at each 1 second period from 5 seconds before the onset of a referring expression to 10 seconds after the onset. Figure 4 shows the difference of the averaged

eye gaze matching rate over each period between the successful and unsuccessful pairs. The label "$t \sim$" denotes a time period of $[t, t+1)$ in second. The successful pairs show a higher matching rate around the onset of a referring expression than the unsuccessful pairs. In addition, the peak of the matching rate comes around 2 seconds after the onset. This tendency supports the result by Richardson and Dale (2005).

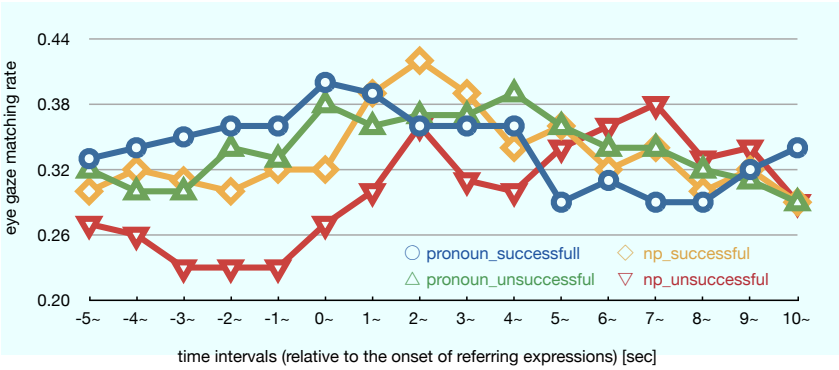


Figure 5. The eye gaze matching rate (pronoun vs. noun phrases).

As described above, each referring expression has various attributes, including its surface form, i.e. pronouns, types of noun phrases and so on. In computational linguistic research, Iida et al. (2010) reported that effective information for automatic reference resolution by computers varies depending on the surface form of referring expressions. Following Iida et al. (2010)'s categorisation, we investigated the tendency of eye gaze matching of pronouns and that of other types of noun phrases separately to find their difference. Figure 5 shows the difference of the averaged eye gaze matching rate between pronouns (331 instances) and noun phrases other than pronouns (359 instances). In both successful and unsuccessful pairs, the peak of the matching rate comes at the onset of referring expressions for pronouns, and at around 2 seconds after the onset for other noun phrases.

The explanation for this observation can be twofold. First the temporal length of pronouns is shorter than that of other noun phrases. In our current data, the average length of pronouns is 0.24 seconds, while that of other noun phrases is 0.76 seconds. In addition, noun phrases generally have more complex syntactic structures than pronouns which usually consist of a single word, therefore noun phrases will require more time to understand their meaning. Second, pronouns tend to be used for salient referents in the preceding context (Grosz et al. (1995)). Thus, it takes little time to gaze at the referent of a pronoun for both participants.

Conclusion

In this research, we recorded eye gaze of both participants in synchronisation with speech in a collaborative problem solving setting, and analysed the data in terms of

eye gaze matching rate. The results showed that (a) the eye gaze matching rate is higher in successful pairs than in unsuccessful pairs, (b) the peak of the matching rate comes at a different position from the onset of referring expressions depending on the surface form of the expressions, i.e. pronouns and other noun phrases.

In this research, we analysed only 8 dialogues by 2 pairs. In order to confirm the above claim, we need to investigate with much larger data. Even with the current data, we investigated eye gaze matching regardless of what they were looking at. We should take into account the target object as well for further precise analyses when the eye gaze matching happened.

Acknowledgments

This research is supported by Grant-in-Aid for Scientific Research (B) (21300049).

References

Brown-Schmidt, S. and M. K. Tanenhaus (2006): 'Watching the eyes when talking about size: An investigation of message formulation and utterance planning'. *Journal of Memory and Language*, vol. 54, no. 4, pp. 592–609.

Clark, H. H. and E. F. Schaefer (1989): 'Contributing to Discourse'. *Cognitive Science*, vol. 13, no. 2, pp. 259–294.

Ehrlich, K. and K. Rayner (1983): 'Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing'. *Journal of Verbal Learning and Verbal Behavior*, vol. 22, no. 1, pp. 75–87.

Grosz, B. J., A. K. Joshi, and S. Weinstein (1995): 'Centering: A Framework for Modeling the Local Coherence of Discourse'. *Computational Linguistics*, vol. 21, no. 2, pp. 203–225.

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

Hanna, J. E. and M. K. Tanenhaus (2004): 'Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements'. *Cognitive Science*, vol. 28, pp. 105–115.

Iida, R., S. Kobayashi, and T. Tokunaga (2010): 'Incorporating Extra-linguistic Information into Reference Resolution in Collaborative Task Dialogue'. In: *Proceedings of 48th Annual Meeting of the Association for Computational Linguistics*. pp. 1259–1267.

Metzing, C. and S. E. Brennan (2003): 'When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions'. *Journal of Memory and Language*, vol. 49, pp. 201–213.

Richardson, D. C. and R. Dale (2005): 'Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension'. *Cognitive Science*, vol. 29, no. 6, pp. 1045–1060.

Spanger, P., M. Yasuhara, R. Iida, T. Tokunaga, A. Terai, and N. Kuriyama (2010): 'REX-J: Japanese referring expression corpus of situated dialogs'. *Language Resources & Evaluation*.

Sturt, P. (2003): 'The time-course of the application of binding constraints in reference resolution'. *Journal of Memory and Language*, vol. 48, no. 3, pp. 542–562.

Tokunaga, T., R. Iida, M. Yasuhara, A. Terai, D. Morris, and A. Belz (2010): 'Construction of bilingual multimodal corpora of referring expressions in collaborative problem solving'. In: *Proceedings of 8th Workshop on Asian Language Resources*. pp. 38–46.

van der Meulen, F. F., A. S. Meyer, and W. J. Levelt (2001): 'Eye movements during the production of nouns and pronouns'. *Memory & Cognition*, vol. 29, no. 3, pp. 512–521.

Zbilut, J. P., A. Giuliani, and C. L. Webber Jr. (1998): 'Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification'. *Physics Letters A*, vol. 246, no. 1-2, pp. 122–128.