Collecting and Annotating Conversational Eye-Gaze Data

Kristiina Jokinen
Doshisha University and University of Helsinki
kristiina.jokinen@helsinki.fi

Seiichi Yamamoto, Masafumi Nishida
Doshisha University, Japan
seyamamo@mail.doshisha.ac.jp
mnishida@mail.doshisha.ac.jp

ABSTRACT
This paper describes our work on the collection and annotation of conversational eye-gaze data. The corpus contains 28 multiparty conversations among participants who are freely chatting with each other on everyday interesting issues, and the corpus includes real-time eye-tracking data of one of the participants. The eye-tracked person is also videoed from the front with a separate camera, and this allows synchronizing the two views of their gaze behaviour. Interactions are balanced so that comparison of communicative behaviour along familiarity, gender, and language between the participants in the different groups is also possible. At the moment, six of the dialogues are annotated, and preliminary analyses on the annotated data are also described.

INTRODUCTION
In about 20 years, technology to track and analyse eye-gaze has evolved to the level where experiments are easy (and cheaper) to arrange, and the resulting data is fairly robust and accurate. Although eye-trackers have long been used as diagnostic tools in medical and cognitive psychology laboratories, their technical evolution has only quite recently allowed a wider range of use for eye-trackers from interface devices to interaction monitoring. Eye-trackers have also become more popular in communication research, and many studies use eye-trackers and include eye-tracking data in order to build models for human-machine interaction, see e.g. Ishii and Nakano (2008) among others. Currently eye-tracker technology is actively developed, and an overview of the technology and the current trends can be found e.g. in Jacob and Karn (2003) and Duchowski (2003).

Our aim in collecting eye-gaze data has been two-fold. First, given the possibility of new technology to collect data on human communicative behaviour, it seems only natural to do so: this allows us to explore the limits of the technology, and simultaneously also to widen the scope of research techniques and methods. Second, since eye-gaze plays a crucial role in fluent communication (Argyle and Cook, 1976), it is important to be able to study the speaker’s gaze behaviour and focus of attention using empirical data also from the signal-level point of view. This provides objective data for interaction studies, and thus complements subjective interpretations of the important communicative events. As we wanted to focus on naturally occurring human communicative activity, we collected conversational corpus which contains data with two distinctive characteristics:

1. It contains three-party conversations instead of two-party dialogues,
2. Conversations exhibit free-flowing associative activity rather than task related activity.

The first aspect has consequences on the participants’ roles and mutual relations, and also on interaction management. In two-party dialogues, the two participants share the space between them and can thus both observe the other and be aware of the other observing them. The participants take turns between the two of them, and both are responsible for dialogue management and for the success of the interaction (given their respective roles). In multi-party dialogues, however, the conversation takes place within a context which is spatially more complex: interaction space is not only larger but it also contains areas which are not directly shared by all the interlocutors: two participants can converse between them, while the others remain observers. This makes interaction management more complex (Healey and Battersby, 2009): some of the interlocutors may be aware of a particular aspect of interaction while the others are unaware of the same aspect. Consequently, also the models for mutual knowledge and grounding of information are more complex, and it can be assumed that the eye-gaze functions as an important signal in interaction management.

Concerning free-flowing vs. task-based dialogues, it was considered useful that communication models be based on data that aims at engaging participants in an activity which is as unconstrained as possible so as to avoid any other additional requirements or cognitive demands on the interlocutors’ behaviour than what was already imposed by the setting as such. Moreover, it can be assumed that any likely differences in the participants’ eye-gaze behaviour would come out most clearly if they could have a friendly chat instead of being required to focus their attention on some (artificial) task. It must be emphasised that this kind of free chatting is not claimed to be more natural or intuitive than interaction in task-related situations; quite

1 The collection and analysis was carried out when the author was the NICT Visiting Scholar at Doshisha University, Japan.
contrary, both are considered natural in what comes to the typical behaviour of the interlocutors, differing only in the activity type and the interlocutors’ social roles. However, as the activity that the speakers are engaged in is known to affect the participants’ behaviour, it can be assumed that free chatting among peers is one of the most neutral types of interaction in this respect: the constraints arise mainly from the communicative needs as such. The results could then be extended to, and compared with situations where more constraints are imposed due to the task (e.g. participants focus their gaze on an object in the shared environment or have very distinctive social roles such as teacher-student, leader-follower).

In this paper we describe our work on the collection and annotation of conversational eye-gaze data. The paper is structured as follows. We first describe the eye-tracker used in the data collection, and then go into details of the data collection setup. We present the collected data, and briefly refer to the annotation work and the results of preliminary analyses. We finish with discussions of future work.

EYE-TRACKER

Usually eye-trackers are desk-mounted video-based systems which have the camera on the desk besides or under the screen, and which can show the focus of the user’s gaze on screen in real time. An additional computer is needed to do the image processing for the eye, although the most advanced systems today integrate the optics for videoing the eyes and the computational processing with the screen. It is also possible to have head-mounted eye-trackers which free the user from sitting in front of a screen.

The actual camera for tracking the eye is shown in Figure 2. The system operates by sending an infrared light beam to the eyes and measuring the angle of reflections from the cornea by the two camera devices. Figure 2 shows the light emitting device in the middle and the two cameras that record the reflection of the light from the right and left eye are on its left and right, respectively.

Figure 2: NAC EMR-AT VOXER Eye-tracker.

Figure 3 depicts the optimal measures needed for calculating the optics of light reflection. The angle between the table and the eye-tracker camera should be about 28 degrees, and the user’s eyes should be at 517mm distance from the eye-tracker camera and about 40 cm higher than the table top on which the camera is placed. In this position, there is about 20 cm margin to move the head forward and backward without disturbing the tracking accuracy.

The tracker also uses the shape of the eye to locate the gaze. The shape is determined in the setup process and calculated with respect to the white and black pattern recognized in the picture of the eyes. Figure 4 shows the control panel for determining eye shapes. The pictures of both eyes by the two cameras are shown so that the person’s right eye is on the left and the left eye on the right. The patterns can be modified by changing the relative amount of black and white parts in them, and thus reach the best fit with the person’s overall eye-shape.

The overall setup and operation of the eye-tracker is managed via the control panel that shows on-line view of the operation of the system. Figure 5 presents a snapshot of
the control panel with a person’s eye being tracked during calibration. It shows the camera view of where the user is looking at (empty table and the calibration board at the back), the two frontal face views by the two camera devices measuring the light reflection, and a close-up view of the eye which is being tracked (in this case the person’s right eye). The reflection points of the beam are also shown as cross points of a horizontal and a vertical line. In the live camera view of the eye in the top right corner of the panel there are two intersecting lines and they correspond to reflections from the pupil and from the cornea of the eye.

Figure 4: Determining the shape of the eye.

During the calibration phase, the user looks at nine points on the calibration board (cf. Fig. 1), while the system measures the eye’s position, shape, and the reflexions of the infrared light. Also head movement is taken into account to compensate movements up, down, and sideways. The sampling rate of the eye-tracker is 60 Hz.

Figure 5: View of the control panel.

DATA COLLECTION
The data was collected during the first author’s visit to the Doshisha University in Kyoto, Japan. The collection setup is shown in Figure 6. Three participants sit in a triangle formation, and one of them (the eye-tracked speaker, ES) has their eye movements recorded by the eye-tracker (the rightmost person in Figure 6). The two other participants, the left-hand speaker (LS) and the right-hand speaker (RS) are videotaped by the camera (in the foreground of Figure 6) as reference points to where ES’s gaze is focused on.

In typical eye-tracking experiments ES looks at the stimulus on a computer screen which is still and stable with respect to the ES. However, our setup differs from this in that we now have a group of three partners who converse with each other and none of them is necessarily still. Quite contrary, instead of a static computer screen, LS and RS are dynamic targets as they can move their head and whole body backward and forward and also tilt their head or bend their body sideways. The optics of the eye-tracker is rather robust, and allows ES head movements of about 20cm depth and 30 degree angle. Thus the participants can have fairly natural body and head movements that are typical for free conversations. However, if the ES head movements are very large, eye-gaze data cannot be captured. Also, if the participants laugh, as they often do during chatting, the eye-tracker loses some data, since ES’s eyes become small and the relevant eye-patterns cannot be found. Of course, when ES blinks, no eye-gaze can be recorded either.

A special problem was caused by the special lamp in front of ES. This was needed in order to help the eye-tracker to distinguish the eye-shape, but it was generally considered rather bad as it shone directly at ES and distracted ES from the two other partners. A better general lighting or spot lights from the ceiling were considered as solutions to the problem in the future collections. In general, the current set up was fine. However, for some participants it did not help in the recognition of their eye pattern: their eyes could not be tracked, apparently due to difficulties in distinguishing between the white and the colour, or to their eye region being too small to be captured by the camera.

Figure 6: Data collection setup

The data collection took place in two phases. In the first phase, which was also a pilot case, six conversations were collected, while in the second phase, 22 conversations were collected with different participants. All conversations are among three interlocutors, and are 10 min long. Before conversation recordings, the participants were told the purpose of the study, and they also signed a consent form that allows video-recordings to be used in research and shown publicly.

As already mentioned, the participants were not involved in any particular task, but were instructed to learn more about their partners and discuss issues that they were interested in. The unfamiliar conversations (see below) were directed especially to elicit conversational information about first encounters and situations where people get introduced to
each other. Consequently, conversations are lively chatting on topics that range from hobbies and weekend plans to studies and travelling. The participants seem to behave naturally despite of being videotaped and the general laboratory conditions. Especially, the restriction of the head movement due to the eye-tracker’s technical limitations did not seem to have a big effect on the naturalness of the dialogues. As an explanation to this, it was also suggested that Japanese people in general move little during conversations and thus the eye-tracker constraint did not have a noticeable effect.

In the first, pilot phase, the participants were six Japanese students from the laboratory (5 male, 1 female). They were familiar with each other although did not necessarily know each other very well. In order to get a mixture of participants with minimum contact with each other in the experimental setting, the participants rotated among themselves so that the eye-tracked person was always a new participant in each triad. The rotation is schematically shown in Figure 7.

Figure 7: Rotation of the participants in the pilot phase

For the second collection we recruited participants from among the students in other laboratories and friends outside the university. There were 18 different participants with their age range in early 20’s, and they were grouped so that we got 14 conversational triads with unfamiliar participants. All subjects were given book cards as tokens of our gratitude.

The instructions were the same as in the pilot case, but it was emphasised that the participants should introduce themselves properly first.

The collection setup was also similar to the one in the first phase, except that we also had a second camera that recorded the eye-tracked person’s face and head movements, see a snapshot in Figure 8. We thus have two types of data concerning the ES behaviour: what the persons themselves focus their gaze on (through the eye-tracker), and how they are seen by the other partners (through the second camera). By synchronizing the two camera views, we will be able to compare the two types, “inside” and “outside” views, of the ES gaze behaviour.

DATA DESCRIPTION

We collected 14 unfamiliar conversations where the interlocutors did not know each other before. Four of the unfamiliar conversations are within female-only groups, four within male-only groups, and six within mixed gender groups (two male students and one female student).

We also collected four familiar conversations among the students of the laboratory, so that together with the six already collected ones the number of familiar conversations would be 10, and the gender balance would equal that of unfamiliar ones (four male-only, six mixed-gender groups). Moreover, four conversations among familiar participants speaking English were also recorded, in order to provide data for comparison between Japanese and English conversations.

Altogether the corpus contains 28 conversations (14 familiar and 14 unfamiliar conversations), balanced with gender, and including eye-tracker information on top of the video data. This amounts to about 280 minutes (4 hours 40mins) of natural conversations, see Table 1.

<table>
<thead>
<tr>
<th>Type</th>
<th>Group</th>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Familiar pilot</td>
<td>Mix</td>
<td>3</td>
<td>Familiar partners – pilot study</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>3</td>
<td>Familiar partners – pilot study</td>
</tr>
<tr>
<td>Familiar</td>
<td>Mix</td>
<td>3</td>
<td>Familiar partners</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>1</td>
<td>Familiar partners</td>
</tr>
<tr>
<td>Familiar English</td>
<td>Mix</td>
<td>4</td>
<td>Familiar English-speaking partners</td>
</tr>
<tr>
<td>Unfamiliar</td>
<td>Mix</td>
<td>6</td>
<td>Unfamiliar partners</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>4</td>
<td>Unfamiliar partners</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>4</td>
<td>Unfamiliar partners</td>
</tr>
<tr>
<td>Eye-tracked face</td>
<td></td>
<td>22</td>
<td>Videos of the eye-tracked person (from all conversations except Familiar-pilot)</td>
</tr>
</tbody>
</table>

Table 1 Statistics of the collected corpora.
An example of the video data is shown in the snapshot in Figure 9. The gaze-path shows that ES shifts focus from left to right, and after some fixations on the forehead and left eye of RS, gaze is focussed on the right eye of RS.

**Figure 9.** Sample video showing a gaze path from left to right.

### ANNOTATION AND ANALYSIS

Five minute clips of each of the six familiar conversations collected in the first phase have also been annotated with dialogue acts, gaze, facial expressions, and turn-taking behaviour, and the annotated data has been used for experimental studies concerning the relation between gaze and turn management. Annotation was done according to the MUMIN annotation scheme (Allwood et al. 2007) which was modified to take the gaze information into account. In addition, the dialogue act annotation was included according to the guidelines developed in the AMI project (www.amiproject.org).

Annotation was done with the Anvil annotation tool (Kipp 2001) by three students who had basic understanding of the task and goals of the exercise, but no previous experience in annotation. The annotator agreement was measured by Cohen’s kappa-coefficient, and reached the kappa value of 0.46. This corresponds to a moderate agreement.

**Figure 10:** Anvil annotation board with a gaze-path on the speaker on the left.

A view of the annotation board is shown in Figure 10. It consists of separate groups for each of the three participants: eye-tracked speaker (ES), left-side speaker (LS), and right-side speaker (RS). LS and RS have tracks for dialogue acts (words), facial display, hand gesturing, and body posture, while ES has tracks only for dialogue acts and gaze behaviour, for the obvious reason that the video only records the ES’ gaze path and voice (the first annotation did not include the frontal video view of ES). However, all annotation levels contain the same features and feature values for communicative functions dealing with feedback and turn management as well as for emotion of the speakers.

We have also conducted several experiments on the annotated data so as to study the relation between eye-gaze and turn-taking. The experiments and results have been reported in our papers (Jokinen et al, 2009; Jokinen et al. 2010). We confirmed earlier observations concerning eye-gaze and its use in interaction management to indicate if the speaker wants to continue talking or is willing to talk. Eye-gaze thus has an important role in smooth communication as it helps the interlocutors to manage their turns in a cooperative manner, and it allows effective interaction management without explicit spoken expressions.

However, we also noticed that in multiparty conversations, the turn management seem to be signalled with head turns rather than eye-gaze, although gaze is important as an initial signal of who could be next speaker. In multiparty conversations, head movement may function as a more visible signal of the speaker’s focus of attention and willingness to take turn. In two-party dialogues, eye-gaze may be enough to signal the partner’s intention to take the turn or to give the turn, but in multi-party dialogues, the participants may not share the context completely and the partner’s focus of attention needs to be expressed in a more visible manner.

### CONCLUSIONS

The collected conversational eye-gaze corpus is one of the few corpora integrating eye-tracker information and, to the best of our knowledge, the first systematic attempt to collect eye-gaze data in natural multi-party conversational setting. The corpus contains conversations among familiar and unfamiliar partners, and also gender balance is taken into consideration by female-only triads. This allows comparison of data along these lines. The conversations are conducted in Japanese and there are also four English conversations by native or near-native speakers, which can provide basis for intercultural comparison on dialogue and gaze behaviour.

The corpus still needs to be annotated and transcribed in a more detailed manner, and the annotation scheme is to be revised as well. However, it is considered a useful start for further activities on analysing eye-gaze and communicative behaviour in natural conversational situations, as well as on collecting and coding multimodal data that includes eye-gaze information. For instance task-based dialogues, as discussed above, would provide interesting extensions and
comparison points to the existing corpus. As for the analysis, we will pursue the two-level approach as advocated e.g. in Jokinen (2009): the data is studied from the point of view of speech and visual signals as well as from the point of view of human interpretation. The signal-level analysis provides empirical evidence of the events that have occurred, while the human annotation assigns meaning to those that are observed as communicatively important events. The study of such complex issues as those related to human communication requires that evidence is collected from different perspectives, and complementing the new multimodal technology with human analysis provides novel possibilities for this. Work is already going on concerning the relation between speech signal and the annotated data, and we can also extend this with techniques that recognize face and body movement. Such work will be useful considering the envisaged future applications that can sense and interact with their environment.

CORPUS AVAILABILITY
The corpus is accessible from the authors, and we invite interested colleagues and researchers to contact us for further collaboration and usage of the data for human-human and human-machine communication studies.

ACKNOWLEDGEMENTS
We would like to thank Kazuaki Harada, Ryosuke Imayoshi, and Shota Sasaki for their help with the experiments and data annotation, and our subjects for taking part in the experiments. The first author would also like to thank NiCT for the grant that allowed her to conduct research in Japan.

REFERENCES