Miraikan SC corpus : A trial for data collection in a semi-open and semi-controlled environment

Miraikan SC Corpus: A Trial for Data Collection in a Semi-open and Semi-Controlled Environment

Mayumi Bono[1&2], Rui Sakaida[1], Ryosaku Makino[3], Ayami Joh[4]
[1] National Institute of Informatics, [2] SOKENDAI (The Graduate University of Advanced Studies), [3] Waseda University, [4]The University of Shiga Prefecture,
[1] 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 JAPAN, [2] Shonan Village, Hayama, Kanagawa 240-0193 JAPAN,
[3] 2-579-15 Mikajima, Tokorozawa-city, Saitama 359-1192 JAPAN,
[4] 2500, Hassaka-cho, Hikone-city, Shiga 522-8533 JAPAN
{bono, lui}@nii.ac.jp, rmakino@aoni.waseda.jp, ayami.joh@gmail.com

Abstract
This paper shows the concept and design of our Miraikan SC Corpus. A well-structured and well-prepared corpus would be useful to engineers for understanding the mechanism of speech production and the nature of social interaction with regard to informing the design of their systems. Applications of the corpus range from speech recognition and dialogue processing to human-agent interaction systems, among others. We started collecting audio-visual data using multiple video cameras and microphones in October 2012 at a science museum in Tokyo, Japan. In this paper, we describe the reason why we chose the museum as a research field for data collection, how we audio-video-recorded the interactions, and how we dealt with personal information in the data set, such as participants' names, jobs, and places of residence.

Keywords: Miraikan SC Corpus, science communication, data collection, publication, personal information

1. Introduction

This paper shows the concept and design of our Miraikan SC Corpus. A well-structured and well-prepared corpus would be useful to engineers for understanding the mechanisms of speech production and the nature of social interaction with regard to informing the design of their systems. Applications of the corpus range from speech recognition and dialogue processing to human-agent interaction systems, among others. We started collecting audio-visual data using multiple video cameras and microphones in October 2012 at a science museum in Tokyo, Japan. In this paper, we describe the reason why we chose the museum as a research field for data collection, how we audio-video-recorded the interactions, and how we dealt with personal information in the data set, such as participants' names, jobs, and places of residence.

2. Concept of Miraikan SC Corpus

The Miraikan Science Communication Corpus (hereafter, Miraikan SC Corpus) is a multimodal interaction corpus composed of numerous face-to-face conversations between science communicators (SCs) and visitors. It was collected at the National Museum of Emerging Science and Innovation (hereafter, Miraikan) in Odaiba, Tokyo, Japan.

2.1 What is Miraikan?

Miraikan is a science museum that was built with the purpose of sharing knowledge, gained from science and technology research, with the public. It has several permanent exhibitions that share current science and technology advances from the following broad bases: human beings, space, innovation, and the information society. The permanent exhibitions are separated into three zones, all of which were produced under the supervision of scientists and engineers working at the forefront of their respective fields. Moreover, it provides three or four special exhibitions per year and some interactive events in which scientists and visitors have the opportunity to communicate directly with each other.

2.2 Who are Science Communicators?

Some science museums in Japan, including Miraikan and the National Museum of Nature and Science in Japan, have trained SCs to share their knowledge of science with people visiting the museum. In 2009, Miraikan started a program to develop SCs. Their official website explains what SCs do as follows:

> "Miraikan trains Science Communicators internally and externally through a unique human development system based on practical science communication activities. At Miraikan, Science Communicators are appointed on a fixed-term system for a maximum of five years. During their terms as Science Communicators, they engage in science communication activities, including providing explanations of exhibits on the exhibition floor, and planning events and exhibitions. At the end of their terms, armed with their experience as Science Communicators, these knowledgeable personnel take up jobs in research institutes, universities, science museums and other museums, corporations, and educational institutions. They also provide training for

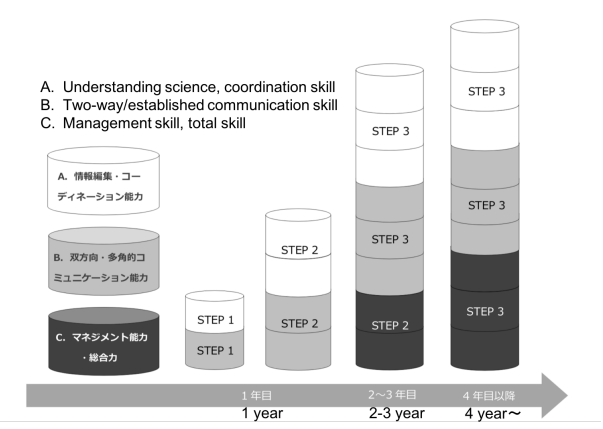

Figure 1: SCs' skills and steps in five years (Miraikan official website)

people who require knowhow in the area of science communication." [1]

Furthermore, they published a report about the purpose and results of the SC program on June 2017, in Japanese (Fig.1)[2].

Miraikan expects SCs to develop three skills each year. The first is the skill of obtaining, understanding, and organizing scientific information as a body of knowledge in their minds. Next is the skill of organizing two-way communication and establishing communication with visitors on the exhibition floors and event spaces. The last skill involves the management and total coordination of SC activities, including teaching and mentoring young SCs.

There is a three-step training process: step 1 is training under senior SCs; step 2 is learning to become an independent decision-maker; and step 3 consists of advanced tasks and mentoring young SCs. This training process has been developing many SCs so far (Table 1).

Table 1: Number of employed persons (2009–2016) (Miraikan official website).

2009	2010	2011	2012	2013	2014	2015	2016	Sum
6	10	10	17	12	13	8	16	92

2.3 Entering the Field

As the Miraikan official website says, almost all of the SCs at Miraikan have five-year employment contracts. They were selected from various academic fields, and they hold masters or doctoral degrees. The SC who explained this arrangement to us also works under this contract, and he had a long academic career before coming to Miraikan. During these five years, SCs learn communication skills aimed at stimulating the public's interest in science and entertaining them. They meet visitors directly on the exhibition floor.

When SCs describe their activities at Miraikan to outsiders, they always say that there is a clear boundary between science communicators and interpreters or curators in a museum. Even if they start by explaining an exhibition to visitors, expert SCs are able to transform this exchange from a one-way explanation to an effective dialogue.

SCs at Miraikan mainly engage in three kinds of activities inside the museum: (1) they explain specific research themes using model kits of objects, such as the brain in a booth where they sit; (2) they sometimes present mini-lectures in a lecture hall, explaining specific topics based on their own experiences and skills; and (3) each of them has time with which to interact with visitors on the exhibition floor, during a pre-determined 1-hour shift.

Starting in October 2012, we conducted a field study at Miraikan in Japan. In April 2012, before we started our fieldwork, one of the science communicators, who had three years' experience as a SC at that time, gave us the opportunity to observe and examine his daily activities at Miraikan. He had been trying to establish criteria with which to evaluate the skills of science communicators and to build an education system for better communication between SCs and visitors (Bono et al. 2014).

2.4 Participants

In general, the term limit of five years at Miraikan is too short to establish and develop scientific knowledge and communication skills on the exhibition floor, especially because the SCs are from diverse academic backgrounds. In addition, they need to plan other activities and events using their specific background in their work places. It is clear that they do not have the time to concentrate too much on brushing up their scientific knowledge and communication skills, to interact with visitors.

Therefore, we decided to make an audio-visual catalog of expert SCs to share methods of interacting with visitors on the exhibition floor at the official kick-off meeting in October 2013. We selected fourteen expert SCs to participate in our recordings and after starting the recordings, other SC recommended one more high-skilled SC, then the total number of SCs for the catalog became fifteen. Half of them had already finished their contracts with Miraikan. We invited them to Miraikan for the recordings. The aim of our project and how we intended to manage the data were explained to all of the SCs and visitors who participated in the recordings, and they granted us permission not only to use the data for our own work but also to publish the video clips, transcripts, and annotation data as a multimodal corpus. This data collection and data publication were allowed by the ethics committee of the National Institute of Informatics.

2.5 Settings and Recording

To construct the Miraikan SC Corpus, we asked Miraikan to allow us to audio-video-record routine conversations between SCs and visitors in the "Spread of Space" and "Challenge the Universe with a Giant Telescope" exhibitions. The reason why we selected these exhibitions was because they offered ideal route for SCs to explain points to visitors; furthermore, it was easy for us and the cameraman to anticipate where SCs and visitors would start their interaction. Additionally, there were high walls, making it possible to avoid filming other visitors who had not agreed to be filmed. The space used for the recordings was separate from other spaces of the exhibition floor.

The shape of the surrounding space resembled a Japanese fan (Fig.2). In many cases, SCs started to explain the exhibition with the planet models from the right edge of the fan. We prepared a rope partition to make a boundary between inside and outside the exhibition and set up a table and chairs to enable visitors to sign the agreement form outside of the exhibition. aJust in case, one of team members was always standing there to receive understandings from other visitors. Next, almost all of the SCs walked through to the Subaru telescope model located at the center of the exhibition. Finally, if visitors had sufficient time, SCs were able to add further explanations of the exhibition using posters and models behind the Subaru telescope. Subsequently, visitors crossed the rope partition placed at the other corner. Afterwards, visitors answered a questionnaire using the table and chairs outside the exhibition.

The recording instruments included six video cameras[3], seven microphones, two lights, and five Kinects (Figure 2) (Trejo et al., 2018). Five video cameras were fixed around the separate area, while a professional camera operator

[1] http://www.miraikan.jst.go.jp/en/aboutus/approach/

[2] http://www.miraikan.jst.go.jp/aboutus/docs/sc_report_201706.pdf

[3] SONY HDR XR550V/PMW EX1R

recorded the front view of the participants with a mobile video camera. Similarly, four shotgun microphones[4] were fixed in position, and several pin microphones[5] were

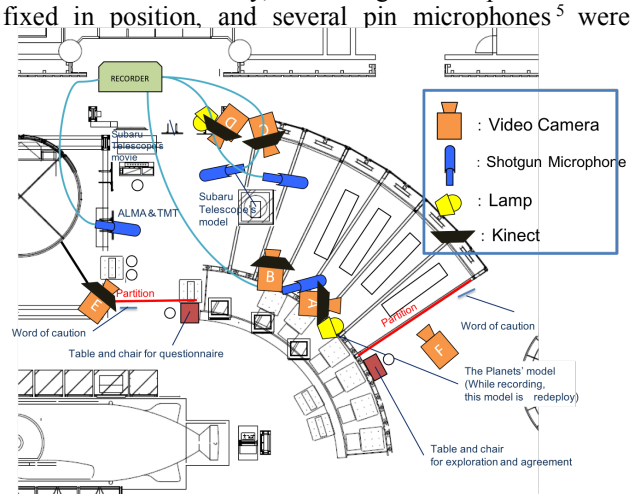

Figure 2: Layout of devices.

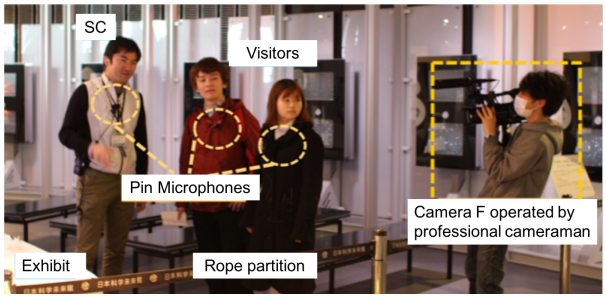

Figure 3: Image of microphones settings and filming by camera F (human operated).

attached to the participants' chests (Figure 3). After the recording, all the recorded sound was mixed, with the noise removed.

3. Data

3.1 Amount of Data

The video recordings were gathered over 10 days in February and March 2014, for about 1 hour per day. On each day, we asked two SCs to talk separately with three groups of visitors, as naturally as possible. Although we did not tell the SCs which exhibit to explain or the route to the next exhibit, they usually followed the same route and explained the same series of exhibits. Most of the SCs started the talk with an explanation of the model of the solar system and invited the visitors to the next exhibit, a model of the Subaru Telescope. Each group consisted of from 1 to 5 people. The average group had 2.26 visitors.

As a result, 15 expert SCs and 79 visitors (44 males, 35 females) participated in the recordings. The participants ranged from 5 to 66 years old, and the average age was 31.26 (*SD* = 15.78). Thirty-five sessions were recorded in total. The total time was 8 hours and 20 minutes. The average length of the sessions was about 14'30". The shortest interaction was around 4 minutes. The longest interaction was around 20 minutes. All of the video and

Figure 4: Screenshot of merged six video clips.

Figure 5: Image of reflection meeting with SCs in backstage.

audio data were merged into one file for each session (Figure 4).

Almost 1.5 hours after the recording, we asked the SCs to have a reflective session, in which researchers (including the authors) and SCs discussed specific video clips backstage for an hour. These sessions were also filmed by one or two video cameras (Figure 5).

3.2 Dealing with Personal Information

One of appealing points of this corpus is that it includes not only speech and text but also video clips to share with the academic community. However, video clips and images always included personal information, especially a participant's face.

Video files

In cases in which video clips included other people who had not agreed to be filmed, we used pixelation to cover their faces. In Figure 6, the upper image includes a person, marked with a yellow dotted square. Because she had not agreed to participate in the filming, we made her unidentifiable through pixelation. On the other hand, the person on the right who is wearing a black top and jeans is a member of this project; thus, no pixelation was needed. The exhibit composed of mirrors sometimes captured human figures; thus, we utilized pixelation, as in the dotted square on the right in the lower image in Figure 6.

Audio files

SCs always explained the exhibit aligning with the story lines predesigned. However, some SCs asked visitors for their name or place of residence, hometown, etc. to

[4] SENNHEISER MKH 416-P48U3/AKG C414-XLII

[5] SONY UWP-V1

Figure 6: Images of masked video clips.

stimulate the visitors' interest and knowledge related to the exhibition as below:

SC: Are you from Tokyo?
V01: umm, no, from *PLACE1*.
SC: Wow, from *PLACE1*, now it's heavily snowing.
V01: Yeah, <laugh>
SC: Where were you born?
V02: *PLACE2*
SC: *PLACE2*, it doesn't have such heavy snow, does it?
V01: No snow.
V02: Yeah, no snow.
SC: Do you know *PLACE3* science museum in *PLACE1*?
V01: uumm, no, I'm sorry.
SC: That museum is…
(1.0)
SC: I'm from *PLACE4*, there is no snow there either.
(1.0)
SC: This (Subaru telescope) is on the top of a mountain in Hawaii.

PLACES 1, 2, and 4 are the names of hometowns for SCs and visitors, which is personal information for them. And PLACE 3 is the name of the science museum in PLACE 1 (V01's hometown). In this case, the SC tried to elicit visitors' interest and uncover their knowledge related to science. If visitors had known the science museum, the SC may have skipped some basic explanations about this exhibition.

We replaced this personal information with a beep sound on the audio file. Furthermore, we replaced place names with a place ID, such as *PLACE1*, *PLACE2* on annotation and transcript files. Our team kept the specific names for certain information- and knowledge-sensitive analyses, such as those involving epistemics in Conversation Analysis (CA) and embodied interaction analysis (Sakaida et al. 2018). We prepared NAME-IDs and JOB-IDs as well. Moreover, we checked the nature of the content in all files. As the results, two files contained politically incorrect topics, so we prepared the other 33 files for publication, with the exception of those two.

4. Publishing Miraikan SC Corpus

We would like to explain our vision for publishing the corpus. We are planning to publish masked video and audio files of interactions on the exhibition floor (except for the reflection meetings backstage), masked transcripts in Japanese on ELAN (.eaf, .pfsx) (Figure 7), and manuals in the near feature. The Miraikan SC Corpus was created in a semi-open and semi-controlled environment, which entails that visitors had an intrinsic motivation to come to the

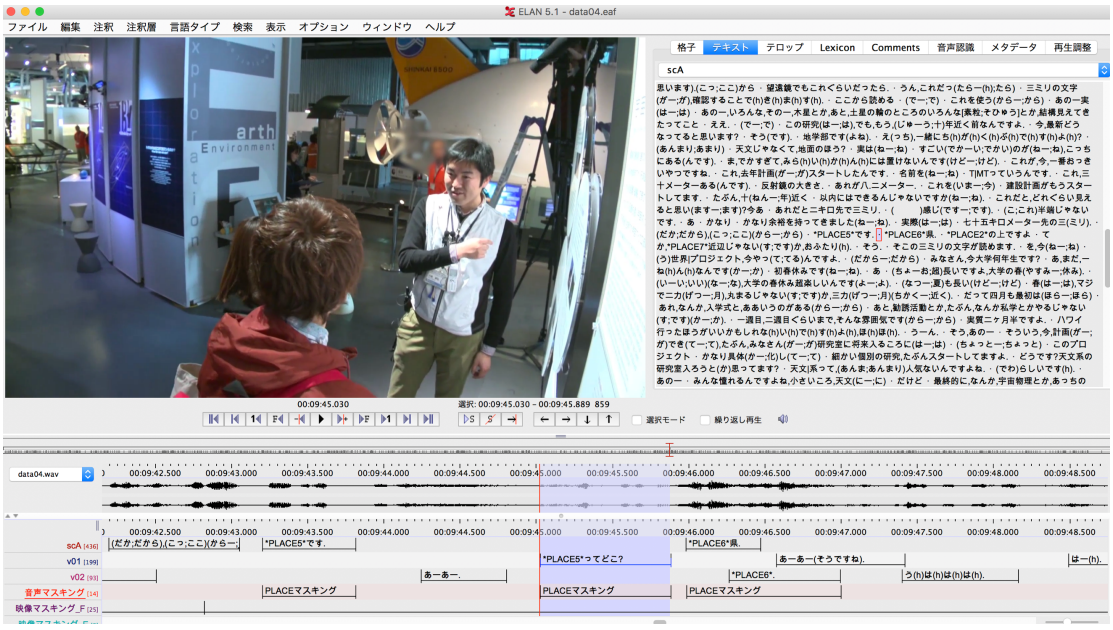

Figure 7: Image of published files on ELAN

Miraikan to engage with the latest results in science and technology. SCs interacted with visitors as part of their daily work, which was filmed.

Previous studies of dialogue corpus research were oriented toward recording high-quality audio data collected in experimental rooms (e.g., the HCRC Map Task Corpus). Using this method, it is difficult to collect naturally-occurring conversation, and it is not a self-motivated interaction. CA analysts have been collecting naturally-occurring conversations for a long time. However, they tend to not release their data, as it contains too much personal information. Furthermore, CA analysts sometimes use this information as ethnographic background in their analyses. Of course, the theoretical backgrounds and research purposes of corpus researchers and CA analysts are very different. However, we assume that some parts of their research interests overlap. In this regard, we tried to create platforms to discuss future directions for interaction studies examining shared and common data sets. Combining two different research fields, such as corpus studies and CA studies, and publishing the Miraikan SC Corpus constitutes a challenge. Furthermore, if some informatics engineers of dialogue and AI systems prove interested in our project, we hope that our efforts may contribute to changing the paradigm of dialogue and interaction studies.

5. Bibliographical References

Bono, M., Ogata, H., Takanashi, K. and Joh, A. (2014). The Practice of Showing 'Who I am': A Multimodal Analysis of Encounters between Science Communicator and Visitors at Science Museum. Organized Session: Brightening Life Style up with Technologies. HCI International 2014, pp. 22-27.

Sakaida, R., Makino, R. & Bono, M. (2018) Preliminary Analysis of Embodied Interactions between Science Communicators and Visitors Based on a Multimodal Corpus of Japanese Conversations in a Science Museum, The 11th edition of the Language Resources and Evaluation Conference (LREC).

Trejo, K., Angulo, C., Satoh, S. & Bono, M. (2018) Towards robots reasoning about group behavior of museum visitors: Leader detection and group tracking. Journal of Ambient Intelligence and Smart Environments, Vol. 10, No. 1, pp. 3-19.