Multimodal Corpus of Multi-Party Conversations in Second Language

Shota YAMASAKI[1], Hirohisa FURUKAWA[1], Masafumi NISHIDA[1], Kristiina JOKINEN[2], and Seiichi YAMAMOTO[1]

[1] Department of Information Systems Design, Doshisha University
1-3 Miyakodani, Tatara, Kyotanabe-shi, Kyoto 610-0321, Japan
[2] Institute of Behavioural Sciences, University of Helsinki
00014 University of Helsinki, Finland
E-mail: dtl0780@mail4.doshisha.ac.jp, dtk0709@ mail4.doshisha.ac.jp, mnishida@mail.doshisha.ac.jp
kristiina.jokinen@helsinki.fi, seyamamo@ mail.doshisha.ac.jp

Abstract

We developed a dialogue-based tutoring system for teaching English to Japanese students and plan to transfer the current software tutoring agent into an embodied robot in the hope that the robot will enrich conversation by allowing more natural interactions in small group learning situations. To enable smooth communication between an intelligent agent and the user, the agent must have realistic models on when to take turns, when to interrupt, and how to catch the partner's attention. For developing the realistic models applicable for computer assisted language learning systems, we also need to consider the differences between the mother tongue and second language that affect communication style. We collected a multimodal corpus of multi-party conversations in English as the second language to investigate the differences in communication styles. We describe our multimodal corpus and explore features of communication style e.g. filled pauses, and non-verbal information, such as eye-gaze, which show different characteristics between the mother tongue and second language.

Keywords: multimodal, second language, turn management

1. Introduction

Language learning is a relevant topic in the cosmopolitan world where it is important to be able to communicate one's message fluently in a foreign language. An effective method to learn a second language is to take interactive language training classes with a private language instructor. Private tutoring, especially tailored for the student's needs, is usually too expensive, and in reality, most students attend classes in which they have to share their teacher's attention. An automatic tutoring system may be used to complement the human instructor with individual training in some cases, such as pronunciation training, and many computer assisted language learning (CALL) systems have been developed and put onto the market especially for these purposes. One of the major problems in commonly used automated language training systems is that the students are assigned a passive role: they are asked to repeat the sentence they had learned or read aloud one of the written choices. As a result, students have no opportunity to practice interaction or construction of utterances on their own.

Advances in speech recognition technologies have pushed forward research on dialogue-based CALL systems, which assign students a more active role by involving them in conversations in which they are able to construct utterances on their own. We also developed a dialogue-based tutoring system for teaching English to Japanese students (Hida, 2012). We plan to further develop this system and transfer the current software tutoring agent into an embodied robot in the hope that the robot will enrich conversation by allowing more natural interactions in small group learning situations.

Human conversations are surprisingly fluent with regard to the interlocutors' turn-taking and feedback-giving behavior. Many studies have shown accurate timing of utterances and pointed out that speakers synchronize or align their behavior to provide robust and efficient communication (Clark & Schaefer, 1989; Pickering & Garrod, 2004; Traum & Heeman, 1997). In the context of interaction technology, especially when considering applications such as robot companions that interact with users in real time, such synchronization is also important. To enable smooth communication between an intelligent agent and user, the agent must have realistic models on when to take turns, when to interrupt, and how to catch the partner's attention. Previous studies (see overviews in (Jokinen et al., 2010a; Jokinen et al., 2010b)) have identified several features that are relevant for interaction control. These include auditory cues such as silent pauses, intonation patterns, and creaky voice; visual cues such as nods, eye-gaze, and mimicry; and language cues such as structural (in)completeness of the utterance, and semantic, pragmatic, and syntactic features. We also need to consider the differences between the mother tongue and second language, which affect the speakers' communication style.

We collected a multimodal corpus of multi-party conversations in English as the second language to investigate such differences in communication styles between Japanese native speakers and Japanese speaking English as the second language. In this paper, we describe the multimodal corpus, and presents some comparison results between the native and second language communication styles. This paper is structured as follows. We first describe the data collection of multi-party conversations in English as the second language. We then present the corpus annotation in Section 3. In Section 4

we compare the features, such as eye- gaze, prosodic features, and dialogue acts in multi-party conversations with those in Japanese as the mother tongue. In Section 5 we discuss preliminary experiments and results on turn management using the corpus and compare the results with those in Japanese. Section 6 draws conclusions and points to future research.

2. Data Collection of Multi-Party Conversations

We have already collected conversational data with eye-tracking information in native Japanese speakers (Jokinen et al., 2010b). These data contain casual interactions in Japanese with Japanese university students who were familiar and unfamiliar with each other. It is used as a reference point to our current data collection setup which focuses on casual English conversations among participants who have English as the second language.

In the current data collections, we collected data from speakers participating in natural, free-flowing conversations, and we also collected data from speakers participating in conversations for achieving a particular goal to compare differences in characteristics between both types of conversations. The corpus is growing and currently contains eight sets of multi-party conversations, each about five minutes long with three participants.

The participants were 16 Japanese university students and a Finn who spoke English at a near-native level. The communicative levels in English of the university students were measured based on the Test of English for International Communication (TOEIC) (toeic, 2012). Their TOEIC scores ranged from 515 to 985 (990 being the highest attainable score). Some participants were persuaded to participate in conversations in Japanese afterwards, and we used their data for detailed feature comparison. Table 1 lists features of the data sets already annotated. Data sets of number 5, 7, and 8 are data of the conversations for achieving a particular goal and the other sets are data of natural, free-flowing conversations.

Fig. 1: Experimental setup for data collection. The caps in the participants' head are the integrated microphone and eye-tracker devices.

Our experiment used groups of three conversational partners who could move according to their conversational activity: they could tilt their heads, gesture with their hands, and bend their body forward, backward, and sideways. The three participants sat in a triangle formation. Their voices were recorded with head-set microphones, and gestures were recorded with three video cameras. The eye-gazes of the participants were tracked using three sets of NAC EMR-9 eye-trackers. Figure 1 shows the experimental setup. Eye-gaze could not always be recorded. This happened when the participant blinked, and in particular when they laughed or their eyes became so small that relevant eye-patterns could not be measured.

set No	participants (TOEIC score)	total speech (sec.)	No of *TG*	No of *TH*
1	Na(735) Ko(670) Mo(690)	258	83	51
2	An(745) Ho(660) Kj(near-native)	252	55	54
3	Hi(910) Sa(780) Kj(near-native)	311	85	36
4	Ka(680) An(745) Kt(985)	300	59	42
5	Ka(680) An(745) Kt(985)	260	70	52
6	Ma(675) Ha(600) Fa(800)	255	74	115
7	Ma(675) Ha(600) Fa(800)	237	61	79
8	Og(515) Oka(590) Oku(600)	158	67	65

Table 1: Features of collected conversational data in English as second language. Abbreviations *TG* and *TH* denote Turn Give and Turn Hold, respectively.

3. Annotations

The corpus is currently growing and being transcribed and analyzed further. The collected data were analyzed at the signal and dialogue levels using the same annotation as the multi-party conversations in the mother tongue (Jokinen, 2010c). The participants' speeches were transcribed by human annotators. On the signal level, the prosodic features of the speeches were analyzed with WaveSurfer (wavesurfer, 2012) and the information concerning the participants' gaze fixation and gaze paths was measured using the eye-trackers. On the dialogue level, an overall spoken dialogue analysis was conducted by manually annotating important dialogue features in the speakers' observed spoken dialogue act. The main goal with the dialogue annotation was to investigate the relation between various communication events, such as dialogue act, eye-gaze, gestures, and prosody, and their communicative functions in turn-management and feedback-giving processes; therefore, the annotations concerned such events and functions. For the dialogue-act annotation, the definitions developed for the Augmented Multi-party Interaction (AMI) project (AMI, 2012) were

followed. For the other annotation features, a modified Nordic Network for MultiModal Interfaces (MUMIN) multimodal annotation scheme (Allwood etal., 2007) was applied.

Turn management refers to the regulation of the interaction flow in conversation with the goal of minimizing overlapping speech and pauses. It is coded with four general features in the annotation: *TurnGive*, *TurnTake*, *TurnHold*, and *TurnNone*. *TurnNone* refers to situations when the partner is listening and has no turn.

Gaze features included a gaze path, coded with the feature *GazeObject*, which refers to the object of the interlocutor's focus. The value *NoGaze* refers to a time span when there is no gaze, either because the interlocutor blinked, laughed, or turned his or her head away from the tracker, which prevented the tracker from recording the gaze. If *NoGaze* was shorter than 0.2 seconds, the gaze elements were regarded as part of the same gaze event (unless there was a gaze shift); otherwise, they were considered separate events, but obviously no shift between them could be recorded. The features and feature values are listed in Table 2.

Annotation features	**Feature values**
Dialogue Act	Backchannel, Stall, Fragment, BePositive, BeNegative, Inform, Suggest-offer, Ask, Other, None
Turn	TurnGive, TurnTake, TurnHold, TurnNone
GazeObject	GazetoRight, GazetoLeft, GazetoOther, NoGaze, None

Table 2: Annotation features applied to experiment

4. Features of Multi-Party Conversations

4.1 Features from Viewpoint of Statistics

Features were compared between conversations in the mother tongue and second language. The annotated data of the Finnish participant was omitted from the analysis described below. Table 3 summarizes the conversation features in English as the second language and Japanese as the mother tongue.

Features	Japanese	English
Av. duration of utterance (sec.)	1.18	1.54
Av. number of filled pauses	5.5	15.8
percentage of *TurnHold* after pause (%)	22	53
percentage of overlapped utterance (%)	24	22
Av. number of eye movements (numbers/min.)	28.4	23.7

Table3: Statistics on multi-party conversation in English and Japanese

The multi-party conversations in English had the following distinguishing features compared with the data obtained in multi-party conversations in Japanese (Furukawa et al., 2011);

(1) Distribution of utterance durations in English was almost the same as that in Japanese, as shown in Fig. 2. However, distribution of utterances shorter than 0.5 sec., which are phrases mainly spoken as supportive responses, was low in English. On the other hand, English conversation had longer utterances than Japanese. As a result, the average duration of utterances in English was longer than that in Japanese.

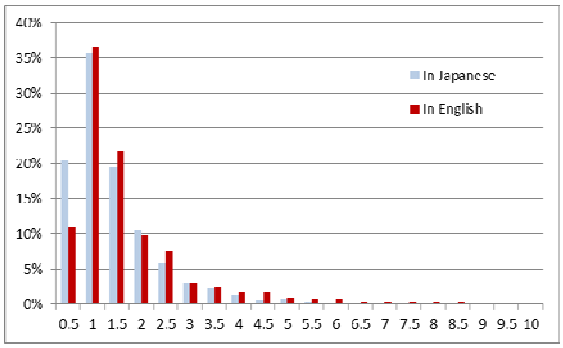

Fig. 2: Distributions of utterance durations in English and Japanese. Figures along the horizontal axis show second-scale utterance durations.

(2) The average number of filled pauses in English was much larger than that in Japanese and the percentage of turn holding after pauses was also larger in English. This suggests that the speakers had difficulties in communicating their message in a single utterance when communicating in the second language.

(3) The percentage of overlapped utterances was almost the same as that in Japanese. This suggests that turn management was similarly coordinated in the native and second language conversations, at least from the viewpoint of keeping overlapping at the same level, considering that the goal with turn management is minimizing overlapping and pauses.

(4) The average number of eye movements, which was measured as the number of eye movements per minute, was smaller in English. This shows that participants observed each other longer to obtain more visual information, although more detailed investigations are necessary to obtain reliable results. (See discussions in Section 4.2).

4.2 Detailed Comparison between Data of Same Participants

To carefully study the differences in the above-mentioned features, we reduced the effects of the other factors such as difference of communication style in participants, and analyzed the conversation data produced in both

languages by from the same participants. The English as the second language data were analyzed from the data sets of 4, 5, 6, and 7 in Table 1, and the Japanese as the native language data were from Japanese conversations from the same participants on the same topics. Figures 3 and 4 show the percentage of duration when the other participants were observing the speaker, and when the speaker was observing the other participants, respectively.

As clearly shown in Figure 3, the percentage of duration when the other participants were observing the speaker in English conversations was longer than that in the native Japanese conversations. This result was true for every participant. On the other hand, the percentage of duration when the speaker was observing the other participants in the English conversations was almost the same as that in the Japanese conversations, although the figures are a little bigger for participants Kt and Fa, who were the most fluent English speakers in each data set. Although the amount of data is not sufficient for drawing statically reliable conclusions, this observation suggests that the participants carefully observed the speaker to understand what he/she was saying, and the speakers with relatively high English proficiency, such as Kt and Fa, were consciously observing the other participants to make sure their messages were delivered and understood exactly as intended.

Fig. 3: Percentage of duration when other participants are observing speaker

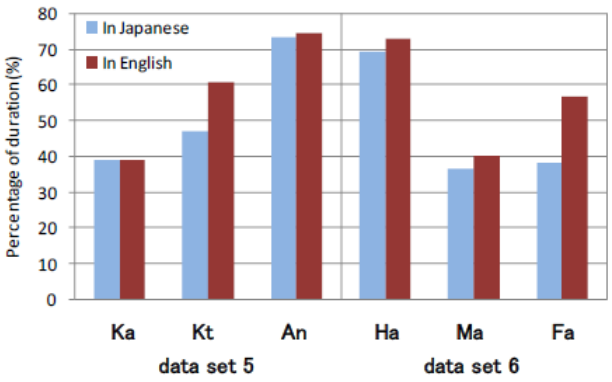

Fig. 4: Percentage of duration when speaker is observing other participants

The above-mentioned experimental results suggest that the differences in these characteristics affect turn management modeling in conversations with English as the second language, especially the feature of eye tracking, which is one of most effective features for turn management in native Japanese conversations.

5. Experiments and Results

To study the effect of the above-mentioned features on a turn management model, we investigated the classification performance of the two most important turn management events *TurnGive* and *TurnHold*. We used the Support Vector Machine (SVM) algorithm, which was used in the classification tasks concerning the conversations in Japanese (Jokinen et al., 2010c). The SVM with parameters trained using the data of multi-party conversations in Japanese was used to classify pauses into *TurnGive* and *TurnHold* using data from both native Japanese and English as the second language. We used the polynomial kernel function and experimentally set the hyper-parameters of SVM.

We then used WaveSurfer to extract the fundamental frequency F0 and speech power (intensity) in the window of 500 msec duration before the end of the utterances, see Figure 5 for a schematic description of the feature extraction and classification. Wavesurfer extracted the F0 and speech power from utterances every 10 msec, and we obtained parameters corresponding to each utterance unit that was based on the end time of the dialogue acts. We used the maximum, minimum, and mean values of the extracted F0 and speech power during the 500-msec window as the speech features for each utterance. We also used the range i.e., the difference between the respective minimum and the maximum values of F0 and speech power, which are used for on-line turn-taking detection (Witten & Frank, 2005).

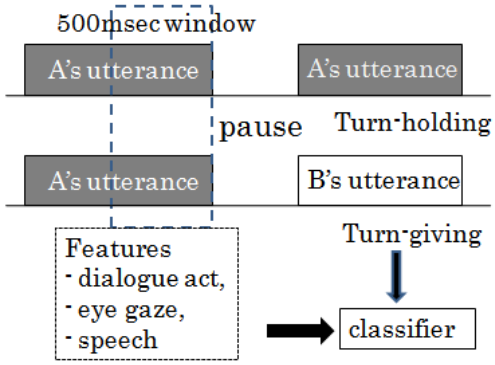

Figure 5: Schematic of turn-giving detection.

Tables 4 (a) and (b) list the classification results obtained for four different feature sets; dialogue act only, dialogue act/eye-gaze, dialogue act/eye-gaze/speech, dialogue act/

speech, in the Japanese and English conversations. The classification performance was high for the Japanese conversations when measuring eye-gaze, as shown by the f-scores for *TurnGive* and *TurnHold* (0.96 and 0.87, respectively). For the conversations in English, however, performance deteriorated, as shown by the f-scores for *TurnGive* and *TurnHold* (0.39 and 0.51, respectively). These results support the difference discussed in Section 4, that eye gaze functions differently in conversations where English is used as the second language compared with conversations where the same speakers speak their native Japanese.

The experimental results in Section 4 also show that the characteristics of eye-gaze depend on the proficiency of the speakers; therefore, we need to consider proficiency as one of the features for the turn management model.

features	Turn-type	prec.	rec.	f-score
Dialogue act (DA)	*TurnGive*	0.75	0.99	0.85
	TurnHold	0.67	0.05	0.09
DA + eye-gaze	*TurnGive*	0.93	0.99	0.96
	TurnHold	0.97	0.78	0.87
DA + eye-gaze + speech	*TurnGive*	0.93	0.99	0.96
	TurnHold	0.97	0.78	0.87
DA + speech	*TurnGive*	0.75	0.99	0.85
	TurnHold	0.67	0.05	0.09

Table 4(a): Classification accuracy evaluated with precision, recall, and f-score for each feature set for conversations in Japanese.

features	Turn-type	prec.	rec.	f-score
Dialogue act (DA)	*TurnGive*	0.65	0.99	0.78
	TurnHold	0.83	0.11	0.12
DA + eye-gaze	*TurnGive*	0.65	0.27	0.39
	TurnHold	0.38	0.76	0.51
DA + eye-gaze + speech	*TurnGive*	0.65	0.27	0.39
	TurnHold	0.38	0.76	0.51
DA + speech	*TurnGive*	0.65	0.99	0.78
	TurnHold	0.82	0.11	0.12

Table 4(b): Classification accuracy evaluated with precision, recall, and f-score for each feature set for conversations in English.

6. Conclusions and Future Plan

We have collected a multimodal corpus for modeling turn management in multi-party conversations where English is used as the second language. We experimented with turn-taking signals, such as dialogue act and eye-gaze, and investigated their effects on turn management in the corpus. The results suggest that turn management in multi-party conversations in English as the second language was different from that in multi-party conversations in Japanese as the mother tongue.

The target users for the dialogue-based CALL system which we are currently developing are Japanese university students. The experimental results suggest that such students' communication styles in English as the second language are different from those in Japanese as the mother tongue, and consequently, we need to develop a separate turn management model for the communications in English as the second language. The quantity of the data collected so far is not sufficient for obtaining reliable statistical models, and we are thus in the process of collecting more data of English conversations under various conditions, in order to develop more reliable turn management models.

Apart from our main research goals, other interesting points to study further concern how the above-mentioned characteristics of the participants who speak English as a second language differ from those of native English speakers, and how the characteristics may change when participants who speak English as a second language communicate with native English speakers. We also plan to collect conversations in both English and Japanese among a bilingual English/Japanese speaker and two native Japanese speakers who speak English as a second language for conducting more detailed analysis.

7. Acknowledgements

This research was supported in part by a contract with MEXT number 22520598. The authors thank to Professor Masuzo Yanagida of Doshisha University for various discussions.

8. References

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management, and Sequencing Phenomena, *Special issue of the International Journal of Language Resources and Evaluation*, 41, 3-4, pp. 273-287, Springer.

AMI (2012): http://www.Amiproject.org/

Clark, H.H., and Schaefer, E.F. (1989). Contribution to Discourse, *Cog. Sci*. 13, pp. 259-294, 1989.

Furukawa, H., Nishida, M., Jokinen, K., and Yamamoto, S. (2011). A multimodal corpus for modeling turn management in multi-party conversations, *2011 Int. Conf. on Speech Database and Assessment* (*Oriental COCOSDA*), pp.142-146, Hsinchu, Taiwan.

Hida, M., Senzai, T., Nagai, Y., Nishida, M., and Yamamoto, S. (2012). Evaluation of a Dialogue-based CALL System, *Proc. 2012 IEICE General Conf.* D-14-4, (2012) (in Japanese).

Jokinen, K., Harada, K., Nishida, M., and Yamamoto, S. (2010a). Turn alignment using eye-gaze and speech in spoken interaction, *Proceedings of INTERSPEECH-2010*, Makuhari Messe, Japan.

Jokinen, K., Nishida, M., and Yamamoto, S. (2010b). On eye-gaze and turn-taking", *In Proceedings of the Workshop on Eye gaze in intelligent human machine*

interaction, (EGIHMI '10). ACM, New York, NY, USA, pp. 118-123.

Jokinen, K., Nishida, M., and Yamamoto, S. (2010c). Collecting and Annotating Conversational Eye-Gaze Data," *Proceedings of the workshop* "Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality" at the Language Resources and Evaluation Conference (LREC), Malta.

Pickering, M., and Garrod, S. (2004). Towards a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, 27, pp. 169-226.

Traum, D.R., and Heeman, P.A. (1997). Utterance units in spoken dialogue," In E. Maier, M. Mast, S. LuperFoy (Eds.), *Dialogue Processing in Spoken Language Systems. Heidelberg: Springer-Verlag*, pp. 125-140.

toeic (2012). http://www.ets.org/toeic.

wavesurfer (2012). http://www.speech.kth.se/wavesurfer/

Witten, I.H., and Frank, F. (2005). Data Mining: Practical machine learning tools and techniques," 2nd Edition, Morgan Kaufmann, San Francisco.