E-025

Joining-in-type Humanoid Robot Assisted Language Learning Systems

カリーファ アルバラ　谷添 友哉　石田充　加藤 恒夫　山本 誠一
AlBara Khalifa[1]　Tomoya Tanizoe[1]　Mitsuru Ishida[1]　Tsuneo Katou[2]　Seiichi Yamamoto[2]

1. Introduction

The rapid progress in transportation systems and information technologies has increased opportunities for worldwide communication, and the ability to communicate in foreign languages is more important than ever. The most effective method of language learning is one-on-one interactive learning with a trained instructor. However, such training is usually too expensive and impractical. In reality, most learners attend classes in which they have to share their teacher's attention with each other. This greatly limits the amount of time each learner spends in producing foreign language speech. Automatic systems have already been used as a complement to human teachers in such areas as pronunciation training.

Thanks to the rapid progress of speech recognition technologies, dialogue-based computer assisted language learning (CALL) systems, which stimulate conversations by encouraging learners to construct utterances on their own and giving corrective responses, have been enjoying a higher profile in recent years [1]-[13]. First- and second-language speech usually have significantly different characteristics such as the phonemes, prosody, lexicon, grammar, dis-fluencies, and so on. Automatic speech recognition (ASR) of second-language speech is still somewhat of a challenge, even for state-of-art ASR. Current speech-interactive language tutors do not let learners create their own utterances because underlying speech recognizers require a high degree of predictability for a reliable performance.

Dialogue-based CALL systems are usually designed to constrain the learners' utterances to obtain high recognition performance for second-language speech. Various methodologies have been proposed for dialogue-based CALL systems to maintain a high degree of predictability by constraining spoken responses by learners. They are divided into two categories. One is called dialogue game, which chats with learners on topics in predetermined domains and gives corrective response to them. The dialogue game is an ideal dialogue-based CALL system if it can recognize a learner's various utterances and generate suitable responses to them. However, second-language speech recognition is still a challenge even for state-of-art ASR systems. As a result, domains of conversations are severely limited in achieving high recognition performance. SCILL [3] covers the topics of weather information and hotel booking. Let's Go [4] is a spoken dialogue system that provides a bus schedule for the area around Pittsburgh, PA, USA. SPELL [5] provides opportunities for learning languages in functional situations such as going to a restaurant, expressing (dis-)likes, etc. The domain of DEAL [6] is trade, specifically a flea market situation.

The other is a method to constrain various expressions by hint stimuli in the form of a keyword or incomplete sentences, or by pre-exercises of typical conversational examples before using CALL systems, and so on [7]-[12]. A CALL application called a "translation game" [7] presents sentences in the learners' native language, asks learners to provide a spoken translation in the target language, and then gives feedback on grammatical and vocabulary errors. This methodology can improve the accuracy of the ASR by reducing the variety of spoken responses compared with conventional spoken dialogue systems. Translation games serve as preparations for dialogue games involving conversational interaction.

Educational use of robots has also been studied in elementary schools, focusing on English language learning [13]. Lee et al. [14] showed that robot assisted language learning (RALL) systems promoted and improved learners' satisfaction, interest, confidence, and motivation.

These CALL systems have been based on ideas that one-on-one interactive learning with a trained instructor is ideal and that dialogue-based CALL systems should have functions of giving explicit instruction to point out the learners' linguistic shortcomings in some way. However, as Lee et al. critically pointed out [14], there is a dearth of empirical research on the developmental benefits of systems with such feedback functions. Most discussions of the publications are largely system descriptions within the dialogue-based CALL literature.

Yamamoto et al. created a multimodal corpus of three-party conversations in their mother tongue (L1) and second language (L2) and compared the behavior of interlocutors in three-party conversations in their L1 and L2. The analyses results showed that they paid more attention to utterances by other participants in conversations in L2 than those in L1 and that a participant of lower L2 proficiency tended to mimic utterances of a participant of higher L2 proficiency [15]. This phenomenon can be regarded as similar to what has been known as interactive alignment in spoken dialogue [16].

On the basis of this observation, we tried an experiment to investigate whether this phenomenon occurs in conversations between a humanoid robot and a human. In this experimental setup, two humanoid robots are simulated as interlocutors of higher proficiency and a human plays the role of an interlocutor with lower L2 proficiency. We also discuss the applicability of this phenomenon to a new speech-interactive language tutoring system, "Joining-in-type Humanoid Robot Assisted Language Learning Systems", on the basis of the experimental results.

This paper is structured as follows. Section 2 reviews related work and introduces the concept of the joining-in-type humanoid

[1] Graduate student at Doshisha University
[2] Professor at Doshisha University

Language Learning Systems, which is based on analyses of second-language conversations. Section 3 illustrates data collection using the joining-in-type humanoid Robot Assisted Language Learning System. Section 4 analyzes the collected speech data. Section 5 discusses our experimental results. Finally, Section 6 presents our conclusion and future work.

2．Related Work

Dialogue-based CALL systems were designed to present explicit instructions from an interaction-partner or instructor to point out learners' linguistic shortcomings. However, second-language acquisition (SLA) requires a balanced learning curriculum that provides opportunities for implicit learning from meaning-based usage and explicit attention to form in use contexts [17].

2.1 Implicit and Explicit Learning

Implicit learning is acquisition of knowledge about the underlying structure of a complex stimulus environment by a spontaneous learning process, simply and without conscious operation. Explicit learning is a more conscious kind of problem-solving where the individual makes and tests hypotheses in searching for structure. Implicit and explicit learning promote different aspects of L2 acquisition (SLA). There is now broad consensus within SLA research that implicit and explicit language learning (a) are different and (b) promote different aspects of language proficiency [17].

An important bulk of language acquisition is implicit learning from usage. This implicit learning from usage allows language users to develop mental representations of language that are optimal given their linguistic experience to date.

Nevertheless, such exposure is not sufficient. Many aspects of a second language are unlearnable, or at best acquired very slowly, from implicit processes alone. Explicit learning and explicit instruction can prompt further development, when an interaction-partner or instructor intentionally brings additional evidence of the linguistic shortcomings to the attention of the learner in some way.

However, these research results of SLA suggest that conventional dialogue-based CALL systems cannot give learners sufficient opportunities to produce L2 speech because of their narrowness in conversation domains.

2.2 Multiparty Conversations

In multiparty conversations, turn taking and interaction obviously cannot be coordinated in a similar way as between two speakers who share the coordination responsibility. Unequal proficiency in L2 may lead to unequal opportunities to participate in such conversations. A multiparty conversation consists of "ratified participants" [18], and participants with lower proficiency might be relegated to "side participant" status regardless of their expertise level in the tasks on which they are collaborating.

Yamamoto et al. collected multimodal three-party conversations in L1 and L2 and compared speech and gazing activities in these conversations from various perspectives [15]. They also compared those activities among three interlocutors of different L2 proficiency. They labeled participants with the highest, second highest, and third highest proficiency Ranks 1, 2, and 3, respectively, in each conversational group. Interesting analyses results were (1) total utterance duration (TDU) and average utterance duration (AUD) are smaller in conversations in L2 than those in L1, and (2) the decreasing ratio of AUD of Rank 2 from conversations in L1 to those in L2 was smaller than those of Ranks 1 and 3. This phenomenon might be a kind of "alignment" [16], and participants of the middle-proficiency group (Rank 2) mediate conversations between participants of high and low proficiency.

2.3 RALL Based on Multiparty Conversation

Educational use of robots has been studied in elementary schools, focusing on English language learning [13]. To identify the effects of a robot in English language learning, the researchers placed a robot in the first and sixth grade classrooms of an elementary school for two weeks and compared the frequency of students' interaction with their English test scores. They showed that the amount of time that children interacted with the robots did significantly and positively affect their English learning. Lee et al. designed a course in which intelligent robots act as sales clerks and showed that, although there was no significant difference in listening skills, speaking skills improved with a large effect size [14]. These robot assisted language learning (RALL) systems promoted and improved learners' satisfaction, interest, confidence, and motivation.

These RALL systems are based on dialogue between a robot and learners. Yamamoto et al. showed that an interlocutor of lower L2 proficiency paid attention to speech of an interlocutor of higher L2 proficiency and that the interlocutor of lower L2 proficiency tended to mimic utterances of the interlocutor of higher L2 proficiency in human conversation. If we can simulate this phenomenon in conversations among two humanoid-robots and a human of low L2 proficiency, we can provide learners a chance to integrate implicit and explicit language learning. To verify the hypothesis and explore the novel language learning methodology "RALL based on multiparty conversation", we collect speech and gazing activity data from conversations among two humanoid robots and a learner.

3. Data Collection

We collected speech and gaze data of 30 participants when they listened to English conversations between two robots and gave English answers to questions from the robots in order to verify the hypothesis that the phenomenon shown by Yamamoto et al. occurs in conversations among humanoid robots and a human.

3.1 Experimental Set-up

Two humanoid robots were set on a table, and a participant sat at the table in a triangular formation to them. We used two Aldebaran NAO humanoid robots. One (R1) played the role of Rank 1, and its speech rate was set to slightly faster than that of the other robot (R2). We used three cameras to take video recordings from different angles to help in the annotation

process. A microphone, which was connected to one of the cameras and attached to the learner's head close to his/her mouth, recorded voices of the learner and the robots. The utterances of the learner were the most important in this case, which is why the microphone was set in that position, although the robots' voices were clear enough in the recordings. A NAC EMR-9 eye tracker recorded learners gazes (Figure 1 and Figure 2) with a viewing angle of 62º and a sampling rate of 60 fps. The eye tracker was set on a cap that the learner had to wear during the experiment. A simple LED-equipped device was set on the table so as to appear in the three video recordings taken by the cameras. This device was used as a signal (operated by using a switch) to help in synchronizing the video recordings. It was also connected to the eye tracker controller to reset the timer of the eye-tracking recording. A computer monitor was set between the robots in front of the learner to show instructions and hints in some cases.

Before the experiment started, the learner was told about the experiment and how to react. The main points of the explanation were as follows: (1) you will have a conversation with two robots; (2) there will be two sessions in which you will be asked different questions to answer; (3) if you cannot answer, a Japanese hint will be displayed for you to translate to answer the question; and (4) you should try to speak clearly and naturally. After the explanation, the eye tracker was calibrated for the learner's gazing to be tracked correctly. Then a simple introduction and instructions (in Japanese) about the experiment were displayed on the computer monitor. The monitor also displayed notifications of the start and end of every session. There were three sessions. The first was an introductory conversation between the two robots without involving the learner. The goal of this session was to help the learner to adapt to the built-in speakers quality of the robots. Then the second and third sessions started (after notifications displayed on the monitor) as parts 1 and 2 of the conversation involving the learner. The participant listened to English conversations between the two robots and answered questions from the robots in English.

The actions of the robots were pre-programmed as groups of fixed alternations of conversational scenarios (which will be detailed in subsection 3.2) and gestures. Three conversational scenarios and four head movement conditions were sequentially selected in every experiment in order to cover all alternations. The four conditions were (1) both robots look toward the learner when talking to him/her, (2 & 3) one looks toward the learner and the other looks at a point between the other robot and the learner, and (4) both robots look at the point between the other robot and the learner. Three hand gestures were used during the conversations in all experiments. A Python program run from a PC was used in this case to control the robots through a local area network. A Java program was used on another PC to receive messages from the main PC to display the instructions and information on the monitor.

Since the robots cannot be autonomous at this stage, due to the lack of high quality speech recognition in L2, the Wizard-of-Oz method was used when dealing with learner's responses. An experimenter who was in the same room of the experiment, was controlling the flow of conversation and deciding whether the learner had answered or not. If the learner made no response, the experimenter could choose to display a hint on the monitor in Japanese to help the learner to answer by just translating the hint into English. Another choice was to command the robot to repeat the question more slowly, or even say the answer. Even if the learner could not answer, the experimenter could choose to continue the conversation as if he/she had answered.

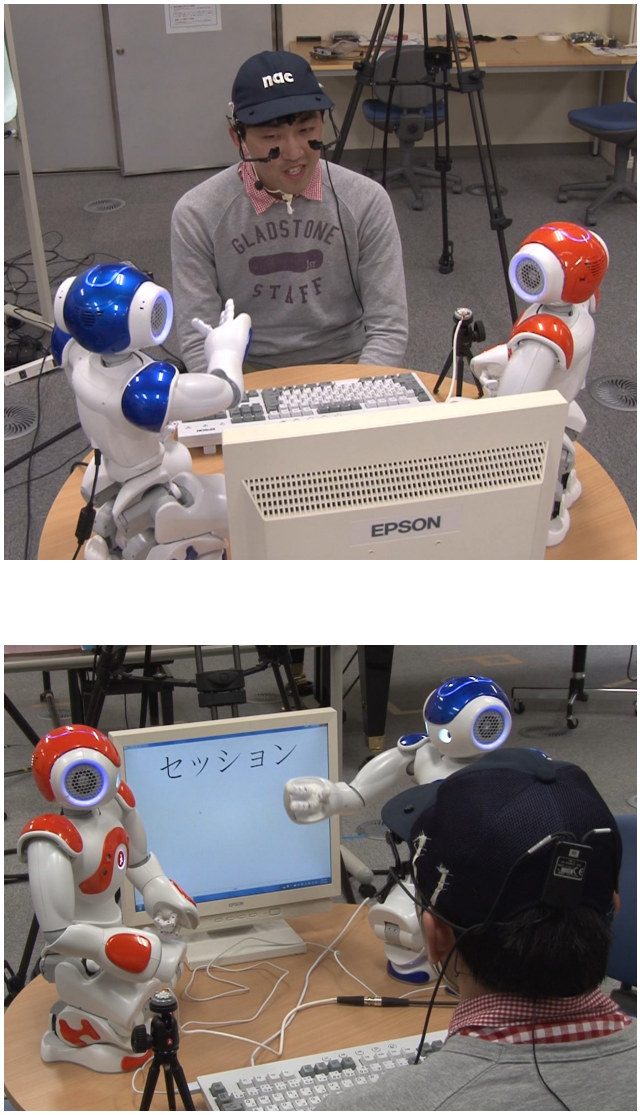

Figure 1: Experimental setup for data collection (Back and front view)

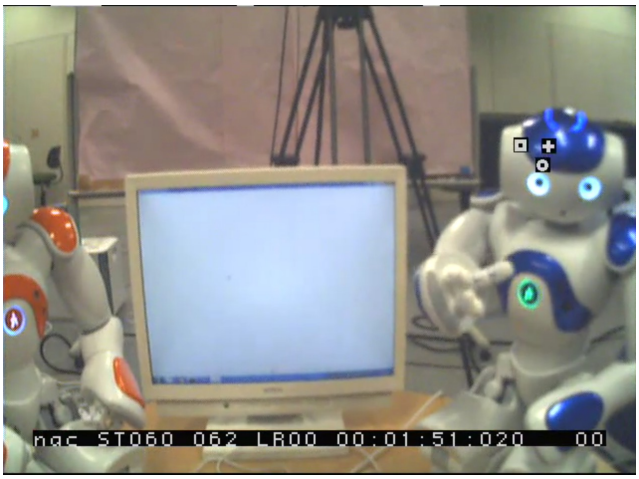

Figure 2: Gaze tracking during data collection

3.2 Conversational Scenario

We created two sets of conversational scenarios. One is a simple conversation of expressing (dis-)likes, and the other was a slightly more complicated one in which two robots collaboratively decide what to take with them on a trip to a mountain. Three variations were constructed from this session to use different wording in every variation. This can help future decisions of choosing proper vocabulary for conversations. In both scenarios, R1 asked the learner the same question that it had asked R2. The design of the conversation was based on topics used in multiparty conversation in the work of Yamamoto et al. [15] and on a similar level of English. The learner was expected to mimic the answer said by R2 or translate the hint displayed (if it was displayed). Below are examples of conversations for both scenarios with examples of the learner's response:

Part 1

R1: Hi
R2: Hi
R1: How are you?
R2: I am fine, thanks

R1: What do you want to take when you go to a mountain?
R2: If I have to go to a mountain, I will take a tent and a knife
R1: How about you? What do you want to take when you go to a mountain?
Learner (example 1): **mountain! ... If I go to the mountain, I will have bring the rope**
Learner (example 2): **I take to mountain... I go to mountain take Knife**
R1: Good answer

Part 2

R1: What is your favorite food?
R2: My favorite food is apples.
R1: How about you? What is your favorite food?
Learner (example): **My favorite food is spaghetti**
R1: Very good, thanks.
R2: Great, thank you.

3.3 Participants

A total of 30 learners between the ages of 18 and 24 were recruited to participate in this experiment as previously mentioned. They were Japanese university students who had acquired Japanese as their L1 and had learned English as their L2. Half the participants had taken the Test of English for International Communication (TOEIC) [19], and the average of their scores is about 553 (990 being the highest attainable score). About half had stayed in an English speaking country for about two weeks on average. Most had spoken English with a non-Japanese speaker. Although there is no precise way to measure their English proficiency, they were assumed to be as proficient as an average Japanese college student in their field. Their answers to the questions during the experiments generally confirmed this assumption.

3.4 Annotation and Transcription

Their utterances were transcribed, and their gazes toward the two robots were annotated. Robots' utterances were also annotated. All annotations were aligned to each other in order to find different information like gazes at a robot during the robot's utterances. Annotations were done manually by (for example) stating the start and end times of gazing at a robot, or the start and end times of the utterance of the learner. To do this, the video recordings and the audio recording had to be gone through in a frame-by-frame manner. The first and second authors did the annotations with help from a colleague in their lab.

For this analysis, we used the EUDICO Linguistic Annotator (ELAN) [20] developed by the Max Planck Institute for Psycholinguistics, which is a linguistic annotation tool for creating text annotations onto video and audio files.

3.5 Questionnaire

We collected questionnaires from the participants after each experiment to examine their attitude to the experimental setup. An English translation of the questionnaire is listed in Appendix I. The questions were modifications of those asked by Yamamoto et al. [15]. Modifications were needed since the participants in this case included robots, not all humans as in [15]. Some questions were about participants' English level and previous experience with robots. Other questions were about their feelings toward the robots and about their experience in the experiment. Google Forms was used to collect the responses of participants.

4. Analyses

We analyzed participant utterances, gazes, and attitudes in interactions in the conversations. We quantitatively analyzed alignment of utterances in Analysis I, gaze data in Analysis II, and questionnaire answers in Analysis III.

4.1 Analysis I: Alignment of Utterances

Alignment between interlocutors is found in various features such as word order and prosody. We analyzed only similarity between transcribed utterances by the participant and utterance by R2 for the same question or by the translation of the displayed hint if it was used using the Levenshtein distances calculated with dynamic programming (DP) and continuous dynamic

programming (CDP). The algorithms were used to find differences in words (not characters), and the results were divided by the total number of words in the two utterances (the participant's and the robot's). Data shows that learners did not need hints most of the time (62.5%). A higher similarity was found in the simple conversation of expressing (dis-)likes than in the more complicated set of conversations (see Table 1). Since the ultimate goal should be using robots only without hints from a monitor (like in the case of the current experiment), calculating the similarity without comparing it with the hint would be helpful for future comparison.

Conversation Feature	*DP*	*CDP*
Word similarity in (dis-)likes conversation	68.6%	80.6%
Word similarity in mountain trip conversation	68.3%	79.4%
Average word similarity in all conversations	68.4%	80%
Average word similarity in all conversations without hints	57%	58.6%

Table 1 Alignment of utterances data using the Levenshtein distances calculated with dynamic programming (DP) and continuous dynamic programming (CDP)

4.2 Analysis II: Gazing Activities

Gazing activities of the participants were analyzed to measure their attention to utterances by the NAO robots. The data of six experiments was omitted in this analysis because of some technical issues with the eye-tracking recording, which required some human judgments in annotations. We used speaker's gazing ratio and listeners' gazing ratio as the criteria to measure their attention. Speaker's gazing ratio is the ratio of all gazing at robots during the utterances of the learner. It can be defined as:

$$\text{Speaker' s Gazing Ratio} = \frac{DLGUL}{DLU}$$

Where DLGUL is the duration of all learner's gazing at R1 during his/her utterances, and DLU is the duration of all learner's utterances.

Listener's gazing ratio is the ratio of all gazing at robots during their utterances. It can be defined as:

$$\text{Listener's Gazing Ratio} = \frac{DLGUR}{DRU}$$

Where DLGUR is the duration of all learner's gazing at R1 or R2 during their utterances, and DRU is the duration of all utterances of R1 or R2. Table 2 compares the averaged results of this work with those obtained in the previous experiment for three-party conversations in L2 [15].

	Conversation with two robots	*Conversation among three-party*
Speaker's gazing ratio	0.28	0.28
Listener's gazing ratio	0.73	0.57

Table 2 Gazing ratio results comparison

4.3 Analysis III: Learners' Attitude

We analyzed the questionnaire and found that most participants enjoyed the activity and thought it was a good way to learn English (see Table 3). They had the impression that R1 had better English proficiency than R2. The appropriateness of the speaking rate of the robots had an average score of about 48%, and they mostly agreed that the robot waited long enough for them to answer. They evaluated their answer and their confidence in their replies to be about 50%. Not much stress was felt during the experiment. Most participants had never dealt with robots before, which may have affected their attitudes.

Questions	*Averages*
Did you enjoy the experiment today?	77.7%
Do you think this is a good way to learn English?	73.7%
Do you think the R1 robot speaks English well?	81.1%
Did you feel pressure from the R1 robot?	43.4%
Do you think the R2 robot speaks English well?	78.3%
Did you feel pressure from the R2 robot?	42.3%
Did the R1 robot speak too fast?	46.9%
Did the R2 robot speak too fast?	49.7%
Did you feel any stress?	31.4%
Do you have any previous experience with robots?	20.8%

Table 3 Some questions from the questionnaire and averages of their answers

5. Discussion

Alignment of utterances between robots and humans can be noticed in the data, where the similarity between the utterance of the learner and R2 indicates a tendency for the learner to mimic the utterance of the robot. This could mean that the system could convey hints to the learner in a natural and indirect way. This kind of stealth education that is based on implicit learning concept can be applied using different conversational scenarios with a wide variety of topics, which can be extended easily from the system.

More dynamic adaptation can be added to the system in accordance with the learner responses and interaction. Using online gaze tracking (we used offline tracking in this experiment) can help in detecting the status of the learner's attention, which can determin future actions of the robots accordingly. The responses from the learner can help in determining the speech rate of the future utterences of the robots, and/or the choice of the level of linguistic difficulty in the conversational scenario.

The noticed alignment in the data might be a sign of the applicability for using the utterances of the robot to enhance the degree of predictability and thus the performance of the speech recognizer in the case of using L2. This means that the speech recognizer can use a language model designed on the basis of this idea.

The use of the hint in Japanese and the clarity of the built-in speakers of the robot are some of the issues that may affect the results in this experiment. Although the hint was not displayed

most of the time, it helped to increase the similarity of utterances. It would be better to have the whole session conducted in English without such hints in Japanese; however, we need to design longer sessions with better choices for vocabulary inorder to avoid using hints. Another possible issue is the relatively low clarity of the built-in speakers of the robots, which may not be the best for second-language learning sessions.

The calculation of the similarity in subsection 4.1 depended on the dynamic programming (DP) algorithm, which may not be the best choice in this case. The alteration of part of speech in the utterance was not considered. This means that the similarity (for example) between "My favorite food is apples" and "My favorite food is oranges" is not 100% in the current case, even though it should be. We used two versions of DP (shown in Table 1) to have their results used for future judgment on a better way of finding the similarity. We should, also, consider the semantic similarity between the two utterances that may be obtained with WordNet or thesaurus.

Since the setup of the experiment was novel, and most participants had never dealt with robots, this could be a reason for having a higher listener's gazing ratio (shown in Table 2) than the previous three-party conversation experiment [15]. Although some of these results might be affected by the novelty of the system setting, they suggest that the proposed system can highly motivate the human learner. Analysis of some of the data was not yet conducted and planned to be done in future, like, the different effects of the four conditions of robot's gazing at learners, the different choices of scenarios' vocabulary, and other questionnaire results.

6. Conclusion

We proposed a novel language learning model, “joining-in-type humanoid Language Learning Systems”. The model is based on observations that an interlocutor tends to produce speech by mimicking utterances from interlocutors of higher L2 proficiency. This was noticed in the calculated similarity between the utterance of the learner and R2. A high percentage of participants’ attention was noticed to be focused on the speaking robot. The system consists of two humanoid robots that are controlled to chat with each other and with a learner. Learners had a positive impression of the experience.

7. References

[1] Eskenazi, M., “An overview of spoken language technology for education”, *Speech Communication*, Vol. 51, Issue. 10, pp. 832-844, 2009.

[2] Kawahara, T. and Minematsu, N. “Computer-assisted language learning (CALL) based on speech technologies.” IEICE Trans. Inf. & Syst. Vol. J96-D, no. 7, pp. 1549-1565 (in Japanese).

[3] Seneff, S., Wang, C., and Zhang, J., “Spoken conversational interaction for language learning”, *STIL/ICALL Symposium*, Venice, Italy, 2004.

[4] Raux, A. and Eskenazi, M., “Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges”, *STIL/ICALL Symposium*, Venice, Italy, 2004.

[5] Morton, H. and Jack, M. A., “Scenario-based spoken interaction with virtual agents”, *Computer Assisted Language Learning*, Vol. 18, no. 3, pp. 171-191, 2005.

[6] Brusk, J., Wik, P., and Hjalmarsson, A., “DEAL: A Serious Game for CALL Practicing Conversational Skill in Trade Domain”, *The proceedings of SlaTE-Workshop on Speech and Language Technology in Education*. Pennsylvania, USA, 2007.

[7] Kweon, O. P., Ito, A., Suzuki, M., and Makino, S, "A grammatical error detection method fordialogue-based CALL system." *Journal of Natural Language Processing*, Vol. 12, no. 4, pp. 137-156, Dec., 2005.

[8] Wang, C. and Seneff, S., "Automatic Assessment of Student Translations for Foreign Language Tutoring." *Proc. Proceedings of NAACL/HLT*, Rochester, NY, pp. 468-475, April. 2007.

[9] Ito, A., Tsutsui, R., Makino, S., and Suzuki, M., "Recognition of English utterances with grammatical and lexical mistakes for dialogue-based CALL system." *Proc. 9th Annual Conf. of ISCA*, Brisbane, Australia, pp. 2819-2822, Sept. 2008.

[10] Wang H. and Kawahara, T., "Effective Prediction of Errors by Non-native Speakers Using decision tree for Speech Recognition-Based CALL system." *Proc. IEICE Japan*, Vol. 92, no. 12, pp. 2462-2468, Dec., 2009.

[11] Rayner, E., Bouillon, P., Tsourakis, N., Gerlach, J., Nakao, Y., and Baur, C.,"A multilingual CALL game based on speech translation." *Proc.Proceeding of LREC*, Valetta, Malta, 2010.http://archive-ouverte.unige.ch/unige:14926.

[12] Nagai, Y., Senzai, T., Yamamoto, S., and Nishida, M., "Sentence Classification with Grammatical Errors and Those Out of Scope of Grammar Assumption for Dialogue-Based CALL Systems." *Proc. TSD, LNCS 7499*, pp. 616-623, Sept. 2012.

[13] Kanda, T., Hirano, T., Eaton, D., and Ishiguro, H., “Interactive robots as social partners and peer tutor for children: A field trial”, *Human-Computer Interaction*, Vol. 19, no. 1, pp. 61-84, 2004.

[14] Lee, S., Noh, H., Lee, J., Lee, K., Lee, G. G., Sagong, S., and Kim, M., “On the effectiveness of Robot-Assisted Language Learning”, *ReCALL*, Vol. 23, no. 1, pp. 25-58, 2011.

[15] Yamamoto, S., Taguchi, K., Ijuin, K., Umata, I., and Nishida, M., “Multimodal corpus of multiparty conversations in L1 and L2 languages and findings obtained from it”, *Language Resources & Evaluation*, DOI 10.1007/s10579-015-9299-2. 2015.

[16] Garrod, S. and Pickering M., “Why is conversation so easy?”, TRENDS in Cognitive Sciences, Vol. 8, no. 1, pp.8-11, 2004.

[17] Ellis, N. C. and Bogart, P. S. H., “Speech and Language Technology in Education: The Perspective from SLA Research and Practice”, *SLaTE*, Parmington, PA, USA, 2007.

[18] Goffman, E, “Replies and responses”, *Language in Society*, Vol. 5, pp.257-313.

[19] TOEIC, http://www.ets.org/toeic.

[20] ELAN,http://www.lat-mpi.eu/tools/elan.

6. Appendix I: Questionnaire

	Previous experience
1	Have you been to an English speaking country?
2	If yes, how long did you stay?
3	Have you talked to a non-Japanese speaker in English before?
4	Have you taken a TOEIC or TOEFL (or any similar) exam?
5	If yes, which?
6	If yes, what was your score?
7	What is your major?
8	Do you have any previous experience with robots?

	Impression of blue robot (R1) *(The answers to the following questions are on a scale of 1 to 7)*
9	Did you like the robot?
10	Did you feel pressure from the robot?
11	Was the robot friendly?
12	Do you think the robot understood you?
13	Did the robot help you to understand the question?
14	Did the robot wait long enough for you to answer?
15	Did the robot speak too fast?
16	How good was the quality of the robot's voice?
17	Do you think the robot speaks English well?

	Impression of orange robot (R2) *(The answers to the following questions are on a scale of 1 to 7)*
18	Did you like the robot?
29	Did you feel pressure from the robot?
20	Was the robot friendly?
21	Did the robot understand you?
22	Did the robot speak too fast?
23	How good was the quality of the robot's voice?
24	Do you think the robot speaks English well?

	Self-Evaluation (overall) *(The answers to the following questions are on a scale of 1 to 7)*
25	How good was your understanding of the conversations?
26	How good were your replies?
27	How confident were you?
28	Did you get nervous when you spoke?

	Self-Evaluation (while listening) *(The answers to the following questions are scale of 1 to 7)*
29	Did you look at the whole of the robot's upper body while listening?
30	Did you look at the whole of the robot's face while listening?
31	Did you look at the robot's eyes while listening?
32	Did you concentrate while listening to the robots?

	Self-Evaluation (while speaking) *(The answers to the following questions are on a scale of 1 to 7)*
33	Did you look at the whole of the robot's upper body while speaking?
34	Did you look at the whole of the robot's face while speaking?
35	Did you look at the robot's eyes while speaking?
36	Did you concentrate on your utterances?

	About The Experiment *(The answers to the following questions are on a scale of 1 to 7)*
37	Did you enjoy the experiment today?
38	Did you feel any stress?
39	Did the Japanese sentences on the display help you?
40	Do you think this is a good way to learn English?
41	Please write any comments you may have.