Measuring Effect of Repetitive Queries and Implicit Learning with Joining-in Type Robot Assisted Language Learning System

AlBara Khalifa, Tsuneo Kato and Seiichi Yamamoto

Doshisha University, Japan

albara.khalifa@gmail.com

Abstract

Computer assisted language learning (CALL) becomes more realistic and motivating for learners through introduction of humanoid robots. A robot assisted language learning (RALL) system is expected to provide an immersive environment for a second language (L2) learner to prepare for real face-to-face communication. We are developing a joining-in type RALL system using two humanoid robots, one playing the role of a teacher and the other playing the role of an advanced peer learner. The interaction between the two robots and learner is designed to smoothly switch between two learning modes, that is, tutoring and implicit learning, for effective language learning. In this paper, we measured the effect ofimplicit learning with repetitive queries quantitatively with 37 learners divided into two groups with and without interaction for implicit learning. The experimental results showed that the repetitive queries of specific grammatical expressions consistently improved the correct use of the expressions, and the improvement was significantly greater when the peer learner robot presented an answer for implicit learning compared with when there was no assistance from the robot.

Index Terms: robot assisted language learning, implicit learning, repetitive training

1. Introduction

Today's globalization has made communication in a second language (L2) an everyday matter for a large number of people. Computer assisted language learning (CALL) is expected to be supplementary for self-learning of L2. In accordance with technical advances in automatic speech recognition (ASR) and natural language processing (NLP), CALL systems are further expected to support various aspects of language learning, such as pronunciation, stress and accents, lexical choices, idioms and expressions, and grammatical rules. To train general oral skills for communicating in L2, various dialogue-based CALL systems have been proposed, such as SCILL [1], SPELL [2], DEAL [3], POMY [4], and DISCO [5].

To make such a dialogue-based CALL system more attractive and realistic, and to let learners prepare for the real face-to-face communication, robot-assisted language learning (RALL) systems have been proposed [6], [7], [8], [9].

A RALL system has a physical presence that a learner must be aware of while responding to the system. Physical presence was reported to be effective in increasing cognitive learning gains [10]. A RALL system introduces nonverbal modalities such as gestures, nodding, and face tracking into the interaction. Those modalities raise the level of experience closer to a real communication, letting the learner imagine reality. Previous studies, most of which were directed toward children, reported that the introduction of robots enhanced learners' interst, motivation, and engagement [8].

In terms of a learning effect, one-on-one tutoring by a skilled instructor is believed to be the best way to learn a L2 [11], [12]. On the other hand, a learner should be exposed to various learning styles like in a classroom. Though a classroom often has the problem of having too many students to give each of them enough chances to commnunicate with the teacher, a student repeats after the teacher, answers questions, receives correction sometimes, and also learns by viewing the interaction between other students and the teacher. Occasionally, students are asked to collaborate on a more complex task and present their thoughts or ideas on the task. Such a combination of tutoring from a teacher and implicit learning from peer learners is considered effective in learning various aspects of communication in L2. An experimental study reported the effect of accompanying an assistive robot with a human teacher in a L2 classroom for children [13]. Children with the teacher and robot learned and retained more vocabulary than children with only the teacher. On the other hand, another study examined how the social behaviours of a social robot affect child second language learning [14]. Though children showed significant improvement between pre- and post- test in both conditions, the difference of high and low verbal availability made no additional gain.

The combination of tutoring from a teacher and implicit learning from peer learners can be simulated with robots in a clear manner. That is, a learner receives tutoring when a robot asks a question or corrects the learner directly, and the learner learns implicitly when a robot similarly interacts with another robot. Giving some examples, the interaction between robots can provide hints to make a response or simply show a model answer. Robots can even entertain or relax the learner to facilitate spontaneous speaking. Furthermore, it is possible to measure the effect of tutoring and implicit learning by automatically evaluating learner responses through the interaction with robots.

In terms of feasibility, effective implicit learning helps not only the learners but also ASR of L2 speech. Generally, recognizing L2 speech is a challenge even for state-of-the-art ASR engines because L2 speech contains various levels of pronunciation, lexical, syntactic, and semantic errors. For these challenges, implicit learning helps regulate learner utterances. The interaction between robots is expected to encourage learners to use expressions similar to those used by robots.

On the basis of this concept, we previously proposed a novel joining-in type RALL system that uses two humanoid robots for Japanese to learn English [15], [16], [17]. One robot plays the role of a teacher and the other plays the role of an advanced peer learner. Though it is not easy to realize truly flexible learning, a typical form of implicit learning is to borrow a useful expression from what a peer learner uses. We designed several scenarios involving two robots and a learner that let the learner listen to the interaction between the robots and learn useful expressions from the interactions. In this paper, we designed

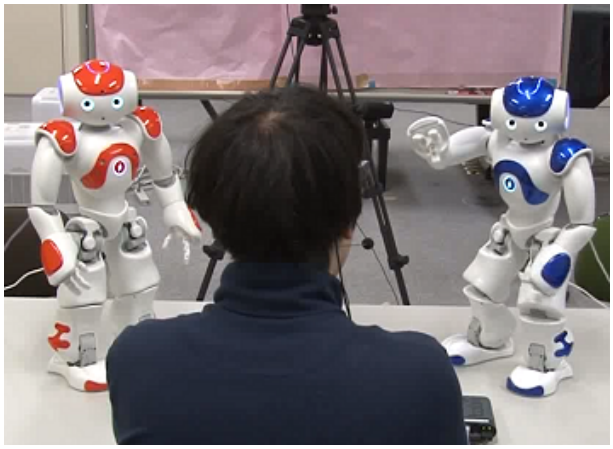

Figure 1: *Snapshot from behind learner.*

a new scenario to measure the effect of implicit learning quantitatively focusing on three English grammatical expressions in conjunction with the effect of repetition, because repetition is essential in measuring the effect of learning [18]. We conducted an experiment with 37 learners to verify the learning effect with the scenario.

2. Joining-in type robot assisted language learning system

2.1. System configuration

The system is configured with two humanoid robots. The robots are set on a table, forming a triangle with a learner sitting at the table. A snapshot is shown in Figure 1. We use two NAO robots. NAO is equiped with the basic funtionalities of speech output, sound production, automatic speech recognition, face tracking, and gesture making. One robot plays the role of a teacher, and the other plays the role of an advanced peer learner. When a robot speaks, it turns its face to whom it is addressing to make it clear for the learner. When a robot speaks to the other robot, the addressed one turns its face to the addressing one, too. Two robots are distinguishable with their voice, and they do not show any visual cues such as light when they speak.

The robots are currently operated in Wizard-of-Oz method because the system is under development and we are collecting a learner corpus on how learners respond to questions and how they behave with the RALL system for further development of a fully automated system. The robots are controlled remotely by an experimenter hidden from the learner through an Ethernet connection. We used recorded speech for speech output from the robots because the scenario is mostly fixed with a limited variation in responses.

2.2. Scenario with focus on specific expressions

After a couple of prototypes [16, 17], we designed a 10-minute long scenario focusing on three specific English grammatical expressions. A part of the dialogue along with the scenario is listed in Table 1. The scenario has a natural dialogue flow to let learners concentrate on the dialogue. A variety of questions that are prompted with the specific English grammatical expressions and expected to be answered with the expressions are asked repeatedly. Specifically, a total of fifteen kinds of questions are asked to the learner with the expectation that the learner will answer them by using three types of expressions that they have been prompted to focus on. That is, five kinds of questions are prepared for each expression. The three expressions are answering of nagative questions, passive voice and a causative verb, all of which are difficult for Japanese learners to use correctly. The five questions for each grammatical expression are listed in Table 2, Table 3 and Table 4. To measure the effect of simple repetitive training, the scenario can be repeated.

Table 1: *Part of scenario involving two robots and learner. (L: learner, R1: teacher robot, R2: peer learner robot)*

speaker	listener	utterance
R1	L	When is your birthday?
L	R1	My birthday is November twenty seven.
R1	L	Ok.
R1	L	What were you given for birthday present last year?
L	R1	An Itunes card.
R1	L	I see.
R1	R2	When is your birthday?
R2	R1	My birthday is May third.
R1	R2	Ok.
R1	R2	What do you think your mother will be given by your father for mother's day?
R2	R1	I think my mother will be given a necklace by my father.
R1	R2	That's great.
R1	L	What do you think your mother will be given by your father for mother's day?
L	R1	I think my mother will be given flower.

The scenario has a flexibility of responses which can be switched depending on the learner's input. For example, the robot is able to repeat the question, or present a sample answer followed with repeating the question if a learner could not respond to the robot, or just pass the current question and continue the scenario in case the learner could not answer at all.

3. Design of experiment for measuring effect of repetitive queries and implicit learning

3.1. Setting of learner groups

To measure the effect of repetitive queries and implicit learning quantitatively, all participants are divided into two groups: a control group and an experimental group. The baseline performance without implicit learning is measured with the control group, and the effect of implicit learning is measured as the difference of the control group and experimental group. Participants of both groups experience the same scenario consisting of fifteen questions but with a different query order.

The order is set as follows. To the control group, the teacher robot always asks a question to a human learner first. To the experimental group, the teacher robot asks a question to the peer learner robot first, and the peer learner robot responds to the question in such a way that the human learner can refer to the

Table 2: *Five negative questions.*

	question
1st	Don't you have a car?
2nd	Don't you like driving?
3rd	Didn't you give a birthday present to your father last year?
4th	Don't you have a video game?
5th	Don't you like science?

Table 3: *Five questions expected to be answered with passive voice.*

	question
1st	What were you given for a birthday present last year?
2nd	What were you taught by your father?
3rd	What were you taught by your mother?
4th	What do you think your mother will be given by your father for mother's day?
5th	What do you think your father will be given by your mother for father's day?

Table 4: *Five questions expected to be answered with causative verb.*

	question
1st	If you have a car and your car is broken, what do you do?
2nd	When you travel, what do you have a guide do?
3rd	When you stay at a hotel, what do you have a receptionist do?
4th	When you visit Tokyo, where do you have a taxi driver take you?
5th	If you employ a private teacher, what do you have a private teacher teach?

Table 5: *Query order of each expression in scenario for control and experimental groups.*

query	control group	experimental group
1st	L first, R2 second	L first, R2 second
2nd-4th	L	R2 first, L second
5th	L first, R2 second	L first, R2 second

L: learner, R2: peer learner robot

expression, but the content itself is not applicable to the human learner's response as it is. Then, the teacher robot asks the question to the human learner second. To compare the performance of the two groups in a fair condition, the first and fifth questions of each expression are asked to the human learner first, and their performance is measured with the questions. The query order for the two groups are summarized in Table 5.

3.2. Two measures for quantifying learning effect

To evaluate the learners' responses in two aspects: correctness in grammatical rules and appropriateness in choice of words and expressions, two measures, expression dependent measure and BLEU score, are computed on the learners' responses.

3.2.1. Expression dependent subjective measure

To measure how many of the participants use a certain expression correctly, we set an expression-dependent criterion for the correct use of each grammatical rule, and judged if each utterance uses the expression correctly or not. For the expressions expected to be used for answering, we set the criteria listed below.

- "Yes" and an affirmative sentence or "No" and a negative sentence for answering of a negative question. Answering with only "Yes" or "No" is not counted as correct.
- A copula and the past participle of a verb in this order for passive voice.
- The causative verb "have", an object and the plain infinitive of a verb in this order for a causative verb.

On the based of the counts of the correct use of the expressions, an expression-dependent ratio of correct use of expression is calculated with equation (1).

$$r_c = \frac{c_{correct}}{c_{total}} \tag{1}$$

where $c_{correct}$ and c_{total} denote the count of correct use and the total count, respectively. This measure shows the ratio of correct use of a certain grammatical rule, but needs a criterion for every grammatical rule.

3.2.2. BLEU score

As a general measure, we evaluate learner utterances with the bilingual evaluation understudy (BLEU) [19] score. BLEU is a popular index for evaluating the quality of machine translation (MT). BLEU score is given by equation (2).

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log p_n\right) \tag{2}$$

where p_n is the precision of *n*-grams in a learner utterance that is determined through comparison with reference sentences. N is usually set at 4, and BP is a brevity penalty, a coefficient for correction. A question from a teacher robot is used as the reference sentence when the teacher robot asks a question to the human learner first, and a sample answer presented by the peer learner robot is used as the reference when the teacher robot asks a question to the peer learner robot first.

BLEU is a general measure, though it is not perfect for evaluating if the learner utterance uses a specific expression correctly or not.

4. Experiments

4.1. Participants and experimental setup

We collected 37 participants between the ages of 18 and 24. They were Japanese university students who had acquired Japanese as their L1 and had learned English as L2. A total of 24 participants out of 37 had taken TOEIC test before. Their scores ranged between 320 and 980, with an average score of 620.

The partipants were divided into the contol and experimental groups, by referencing their TOEIC scores as a counterbalance. Eighteen participants were assigned to the control group, and the other nineteen were assigned to the experimental group. The average TOEIC score of the control group was 616.8, and

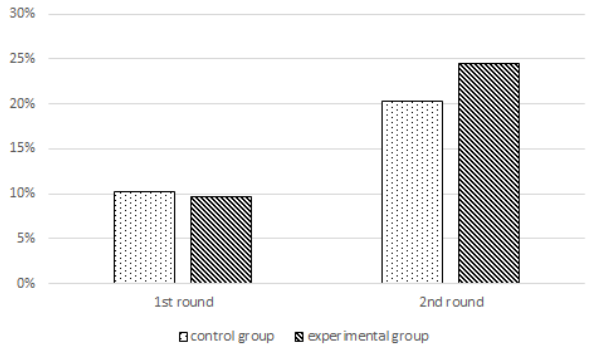

Figure 2: *Overall ratio of correct use of expressions r_c of each round for control and experimental groups.*

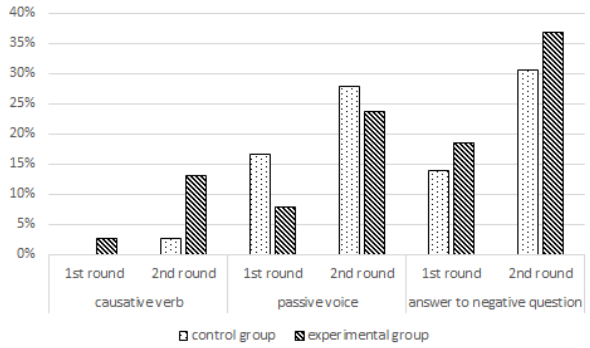

Figure 3: *Ratio of correct use r_c of each expression and round for control and experimental groups.*

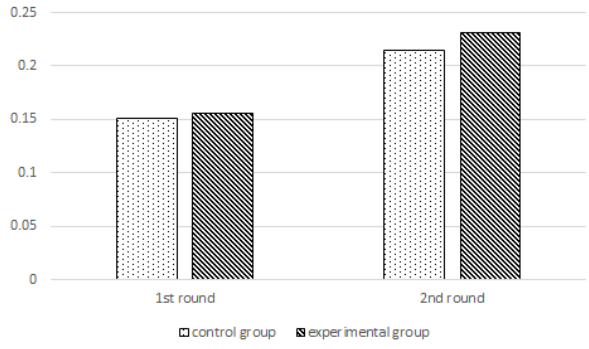

Figure 4: *Overall BLEU score of each round for control and experimental groups.*

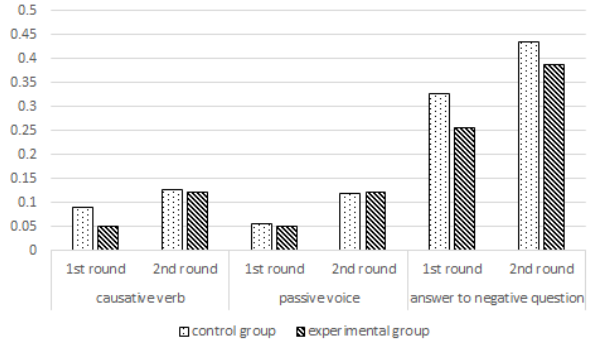

Figure 5: *BLEU of each expression and round for control and experimental groups.*

the average of the experimental group was 622.5. Though the averages did not cover all the participants, the general level of English skill is expected to be comparable between the groups as far as we see from the TOEIC scores available.

The experiment was conducted at a laboratory with recording equipment. Two video cameras placed in front of and behind the learner captured video. Audio was recorded with a headset microphone on the learner as well as the microphones on the video cameras. Information on a learner's point of gaze was captured with a glasses-type eye-tracking system for future analysis.

After wearing the headset microphone and the glasses-type eye-tracking system on and receiving a brief instruction on how to respond to the robots, every participant performed the 10-minute long scenario. To measure the effect of simply repeating the scenario, every participant was asked to repeat the scenario twice with a 5-minute break in between. Hence, every participant was asked to respond to the system with a specific expression ten times, and the performance was measured in the first and fifth questions of the 1st and 2nd rounds.

4.2. Basic statistics of collected data

Every participant had 19 chances to respond to the robots including the 15 questions that were expected to be answered with the expressions that had been prompted to focus on each round. A total of 1,253 utterances were collected from the 37 participants. The average number of words in a learner utterance was 4.5. The total size of vocabulary was 691. The average time required for completing one round of the scenario was 10 minutes.

4.3. Analysis on effect of repetitive queries and implicit learning

Figure 2 shows the overall ratios of using the target expressions correctly for the 1st and 2nd round of the control group and experimental group. The effect of repetitive queries was obvious for both groups, and the degree of improvement was greater for the experimental group than for the control group. The improvement for the experimental group was 15 points while that for control group was 10 points.

Figure 3 shows the ratios of correct use of the causative verb, passive voice, and answering of negative questions. The causative verb was not practical to use orally. Some of the participants seemed confused between the causative verb and the perfect tense of "have". The ratio of correct use stayed at a low level even after the repetitive queries, but the ratio for the experimental group showed a small improvement in the 2nd round. The passive voice of verbs was also difficult for the participants to use correctly. The ratios in the 1st round had a gap between the control group (17%) and experimental group (8%), but the improvement was 9 points for the control group while that was 16 points for the experimental group in the 2nd round. Answering negative questions was the easiest among the three expressions. Though the ratio of correct use was around 15% in the 1st round, it improved soon after repetitive queries. The improvement was 17 points for the control group and 18 points

for the experimental group. Though the correct ratio itself is different between the expressions, every expression showed improvement in the 2nd round compared with the 1st round, and the improvement was greater for the experimental group.

Figure 4 shows the overall BLEU scores for the 1st and 2nd rounds of the control group and experimental group. The scores improved in the 2nd round for both groups, and the improvement was greater for the experimental group. Figure 5 shows the BLEU scores of the causative verb, passive voice, and answering of nagative questions. The magnitude relation of the scores of the two groups was not always equal to that of the ratios of correct use shown in Figure 3, but the effect of repetitive queries was obvious, and the increase was greater for the experimental group compared with the control group.

5. Conclusions

We prototyped a joining-in type robot assisted language learning (RALL) system using two humanoid robots operated in Wizard-of-Oz method. We designed an extended scenario in which learners were asked similar questions expected to be answered with specific expression forms a number of times, and measured the effect of repetitive queries and implicit learning quantitatively with 37 participants collected. Specifically, we divided participants into two groups: a control group without implicit learning and an experimental group with implicit learning, and evaluated their answers with an expression-dependent measure and BLEU score. Both improved consistently as repetitive queries were made, but the improvement was greater in the case where learners responded to the system after repeating answering similar questions with hearing sample answers by the peer learner robot.

The next step is to verify the effect of repetitive queries and implicit learning on memory over a longer term. Furthermore, after collecting enough of a learner corpus to train a language model for the L2 learner speech, we will develop a fully automated system based on ASR. We will attempt to automatically generate dialogue scenarios from a variety of topics in CogInfoCom [15].

6. Acknowledgements

This research is supported by a contract with MEXT, number 15K02738.

7. References

[1] S. Seneff, C. Wang, and J. Zhang, "Spoken conversational interaction for language learning," *Proc. of STILLICALL Symposium 2004*, pp. 3111–3119, 2004.

[2] H. Morton and M. A. Jack, "Scenario-based spoken interaction with virtual agents," *Computer Assisted Language Learning*, vol. 18, no. 3, pp. 1532–1543, 2005.

[3] J. Brusk *et al.*, "Deal: A serious game for call practicing conversational skill in trade domain," *Proc. of SLaTE 2007*, 2007.

[4] K. Lee *et al.*, "Postech immersive english study (pomy): Dialog-based language learning game," *IEICE Transactions on Information & Systems*, vol. Vol.E97-D, no. 7, 2014.

[5] J. V. Doremalen *et al.*, "Evaluating automatic speech recognition-based language learning systems: a case study," *Computer Assisted Language Learning*, vol. 29, no. 4, pp. 833–851, 2016.

[6] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: a field trial," *Human Computer Interaction*, vol. 19, no. 1, pp. 61–84, 2004.

[7] J. Han, "Emerging technologies: Robot assisted language learning," *Language Learning and Technology*, vol. 16, no. 3, pp. 1–9, 2012.

[8] S. Lee *et al.*, "On the effectiveness of robot-assisted language learning," *ReCALL*, vol. 23, no. 1, pp. 25–58, 2011.

[9] T. Belpaeme *et al.*, "L2tor - second language tutoring using social robots," *Proc. of Int. Workshop on Educational Robots*, 2015.

[10] D. Leyzberg *et al.*, "The physical presence of a robot tutor increases cognitive learning gains," *Proc. of Cognitive Science 2015*, pp. 1882–1887, 2012.

[11] M. H. Long, "Input, interaction, and second-language acquisition," *Annals of the New York Academy of Sciences*, vol. 379, no. 1, pp. 259–278, 1981.

[12] B. S. Bloom, "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring," *Educational researcher*, pp. 4–16, 1984.

[13] M. Alemi *et al.*, "Employing humanoid robots for teaching english language in iranian junior high-schools," *Int. Journal of Humanoid Robotics*, vol. 11, no. 3, pp. 4–16, 2014.

[14] J. Kennedy, P. Baxter, E. Senft, and T. Belpaeme, "Social robot tutoring fir child second language learning," *Proc. of ACM/IEEE Int. Conference on Human-Robot Interaction (HRI) 2016*, pp. 231–238, 2016.

[15] G. Wilcock and S. Yamamoto, "Towards computer assisted language learning with robots, wikipedia and coginfocom," *Proc. of IEEE Conf. on Cognitive Infocommunications*, pp. 115–119, 2015.

[16] A. Khalifa, T. Kato, and S. Yamamoto, "Joining-in-type humanoid robot assisted language learning system," *Proc. of LREC 2016*, pp. 245–249, 2016.

[17] M. Ishida, A. Khalifa, T. Kato, and S. Yamamoto, "Features of learner corpus collected with joining-in type robot assisted language learning system," *Proc. of Oriental COCOSDA 2016*, pp. 128–131, 2016.

[18] P. Baxter, J. Kennedy, E. Ashurst, and T. Belpaeme, "The effect of repeating tasks on performance levels in mediated child-robot interactions," *Proc. of Robots 4 Education Workshop 2016*, 2016.

[19] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation," *Proc. of the 40th annual meeting on association for computational linguistics*, pp. 381–388, 2002.