International Journal of Social Robotics (2024) 16:743–761
https://doi.org/10.1007/s12369-024-01134-0

Comparison of Outcomes Between Robot-Assisted Language Learning System and Human Tutors: Focusing on Speaking Ability

Takamasa Iio[1] · Yuichiro Yoshikawa[2] · Kohei Ogawa[3] · Hiroshi Ishiguro[2]

Accepted: 19 March 2024 / Published online: 11 April 2024

Abstract

This study explores how much current mainstream Robot-Assisted Language Learning (RALL) systems produce outcomes compared to human tutors instructing a typical English conversation lesson. To this end, an experiment was conducted with 26 participants divided in RALL (14 participants) and human tutor (12 participants) groups. All participants took a pre-test on the first day, followed by 30 min of study per day for 7 days, and 3 post-tests on the last day. The test results indicated that the RALL group considerably improved lexical/grammatical error rates and fluency of speech compared to that for the human tutor group. The other characteristics, such as rhythm, pronunciation, complexity, and task achievement of speech did not indicate any differences between the groups. The results suggested that exercises with the RALL system enabled participants to commit the learned expressions to memory, whereas those with human tutors emphasized on communication with the participants. This study demonstrated the benefits of using RALL systems that can work well in lessons that human tutors find hard to teach.

Keywords Robot-assisted language learning · L2 learning · Human–robot interaction · Comparison between humans and robots

1 Introduction

Robot-assisted language learning (RALL) has been actively studied over the past two decades. Robots are now being expected to provide students with high-quality education through face-to-face interactions [6, 21, 37, 43]. Compared to video-displayed agents, physical robots can yield more compliance to their requests [6], elicit social behavior from learners beneficial to learning [22], be more engaging and enjoyable [26, 27], and be perceived positively [31, 36, 45]. These findings suggest that robots can generate positive perceptions and improve task performance in educational situations because of their physical presence.

Therefore, this study focuses on a robot playing the role of a tutor in second language learning. In RALL studies, the robot plays several roles: teacher or teacher's assistant in classrooms [1, 2, 8, 16, 17, 49, 51], tutor teaching via one-to-one interactions [13, 14, 18, 23, 29, 39], peer learning with students [5, 20, 33, 52], or novice instructed by students [35, 40, 41]. Although these classifications are not strict and sometimes ambiguous, tutoring roles have been adopted in past studies [6, 37].

One interesting question regarding RALL systems is the extent to which they can improve the students' skills compared to that with a human tutor. The development of large language models has opened up the possibility for RALL systems which are able to teach as flexibly as human tutors.

Yuichiro Yoshikawa, Kohei Ogawa and Hiroshi Ishiguro have contributed equally to this work.

✉ Takamasa Iio
tiio@mail.doshisha.ac.jp

Yuichiro Yoshikawa
yoshikawa@irl.sys.es.osaka-u.ac.jp

Kohei Ogawa
k-ogawa@nuee.nagoya-u.ac.jp

Hiroshi Ishiguro
ishiguro@sys.es.osaka-u.ac.jp

[1] Faculty of Culture and Information Science, Doshisha University, 1-3, Tatara-miyakodani, Kyotanabe-city 610-0394, Kyoto, Japan

[2] Graduate School of Engineering Science, Osaka University, 1-3, Machikaneyama-cho, Toyonaka-city 560-8531, Osaka, Japan

[3] Graduate School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-city 464-8601, Aichi, Japan

However, even with LLM, it is still difficult for robots to progress through lessons, responding freely to students' questions. This study focuses on what the robot can certainly do with the current technology. Specifically, the robot acts as a tutor who role-plays a predetermined scenario, and the scenario is proceeded by the learner's input to a tablet. Such a system integrating a robot and a tablet is recommended in the guidelines for designing social robots as second-language tutors [7]. Despite such limited lessons, it is meaningful to compare the learning outcomes when using RALL systems and human tutors. If the RALL systems do not match human tutors, analyzing why they are insufficient can help improve their functionalities. Conversely, if RALL systems are found comparable to human tutors, this finding can provide evidence regarding the value of RALL systems, which can help deploy RALL systems in homes and schools.

However, to the best of the author's knowledge, direct comparisons between RALL systems and human tutors are limited, not only for language learning but also for educational social robots. A recent comprehensive review of social robots in education [6] states, "Many studies using robots do not consider learning in comparison with an alternative, such as computer-based or human tutoring, but instead against other versions of the same robot with different behaviors... Comparisons between robots and humans are rare in the literature, so no meta-analysis data were available to compare the cognitive learning effect size". Few RALL studies [10, 32, 35, 42, 47, 48] employed experiments and surveys that compared robots and humans; however, they did not achieve clear results, i.e., they did not have statistical tests, lacked statistical information, and did noten robots and humans are rare in compare learning outcomes.

This study explores how much current mainstream Robot-Assisted Language Learning (RALL) systems produce outcomes compared to human tutors instructing a typical English conversation lesson. Among the four abilities in second-language learning (reading, writing, listening, and speaking), speaking is essential for social interaction with others, along with listening. Given that speaking skills are developed through interaction with others, it is appropriate to compare the learning outcomes between the robot and human tutors. Therefore, we focused on speaking skills in second-language learning. Further, we conducted online lessons, in the style of lessons taught by human tutors. Online lessons are less expensive than private lessons at English conversation schools, and they allow students to take lessons at home. Online lessons are becoming increasingly common and are now a popular style of second-language learning. Thus, it is an appropriate subject for comparison with lessons provided by a robot tutor.

This is an exploratory study. We do not make specific hypotheses about the learning outcomes of robot and human tutors because there is little scientific evidence to support such hypotheses. A meta-analysis of the effectiveness of human tutors and intelligent tutoring systems (ITS) in STEM education [44] indicated that the effect size for human tutors was 0.79, and the ITS effect size was 0.76. Belpaeme et al. reviewed educational robots and reported that the effect size for RALL systems was 0.70 [6]. These results confirm that the effect sizes of the human and RALL systems may be comparable. However, the interactivity and flexibility of human tutors in their teaching style may have a significant impact on performance because such features of human tutors are considered important when attempting to improve speaking skills in second language learning. Therefore, it is difficult to formulate a hypothesis solely based on current knowledge.

The contributions of this study is to provide novel and reliable data that can be useful in RALL research and HRI in education, as well as insights into the suitability of robots in education based on that data. Specifically, the following points illustrate the value of our study:

- Few previous studies have compared between robot and human tutors.
- Our data are derived from a relatively long period of work (30 min per day for 7 days) and appropriately reflect the tutors' instruction.
- We compared the RALL systems against competent human tutors recruited from a language school.
- We provide detailed analysis of changes in speaking skills based on previous literature.
- Based on that detailed analysis, we discuss the suitability of robot and human tutors in tutoring.

This paper is organized as follows. First, Sect. 2 briefly describes the RALL study and shows that there has been insufficient comparison between robot and human tutors in the field of education as well as RALL. Next, Sect. 3 describes the methodology of this study. In this section, in addition to the experimental design, we present the details of the learning instruction provided by the robot and human tutors. Then, in Sect. 4, we present the data on participants' speaking skills obtained through the experiment. Section 5 discusses the interpretation, implications, and scope of the data. Finally, Sect. 6 presents the conclusions of this study.

2 Related Work

2.1 Overview of RALL

RALL is a field of educational robotics that targets first- and second-language learning. The use of robots for language learning offers several advantages over existing technologies. For example, Belpaeme et al. [6] makes three points relevant not only to language learning but also to education. They stated, "(i) they can be used for curricula or populations that require engagement with the physical world, (ii) users show more social behaviors that are beneficial for learning when engaging with a physically embodied system, and (iii) users show increased learning gains when interacting with physically embodied systems over virtual agents". These points are strongly related to the embodiment of robots.

Robots do have advantages over traditional computer-assisted instruction because of their bodies, which enable interaction with others and the environment. Interaction between students and their physical environment is important for human language development [4, 15, 19, 46]. Social interactions with others may be important for language development and learning [28]. Many HRI studies have shown that robots can interact naturally with humans through their physicality. The manipulation of real objects [25] and the use of body movements and gestures [34, 38] help children acquire vocabulary. Robots can manipulate objects and use gestures/language as a tool to interact with others. In addition, a systematic review of robot gestures suggests that robot gestures may have positive educational effects [9]. For example, interactions with a gesturing robot were rated more positively and found to be more enjoyable, and gestures have also been shown to help maintain engagement during ongoing interactions. Furthermore, a robot that uses gestures is perceived as a better facilitator of learning.

2.2 Comparison Between Human Tutors and RALL Systems

Only a few studies compared RALL systems and human tutors in RALL. These studies compared robot and adult tutors in children's L1 learning [47, 48], robot and child peers in children's L2 learning [35, 42], and teleoperated robot and human facilitators in L2 learning for adults [32]. In addition, one study conducted an exploratory analysis to compare the robot and human tutors [10]. However, they did not find significant differences in learning outcomes between the robot and human tutors.

Westlund et al. [47] compared conditions in which children learned new words using only a tablet, a tablet together with a robot, and a tablet with an adult experimenter. The results showed no significant differences in the number of nouns learned by the children (means, statistics, and effect sizes for each condition are not stated). Westlund et al. [48] compared the conditions of learning with a robot and an adult in a similar task. This experiment also showed no significant differences in the children's recall (effect size not stated).

Mazzoni et al. [35] used a robot as a peer (novice) in children's L2 vocabulary learning. They compared conditions under which the children learned from each other and from the robot. They reported that children learned more words in the robot learning condition than in the child learning condition. However, the sample size was small, and there was no test for significance between conditions. Van den Berghe et al. [42] used robots as peers in children's L2 vocabulary learning. The results of the experiment showed that there were no significant differences between learning with the robot and learning with the child with respect to vocabulary translation and comprehension.

Lopes et al. [32] conducted a field trial to compare conversations with a teleoperated robot facilitator and a human facilitator in a language café. People rated conversations with the human facilitator as superior in almost all respects. However, this study compared the users' impressions of each conversation rather than the learning outcomes. The robots evaluated in the field study of Lopes et al. were remotely controlled and were not autonomous.

Demir-Lira et al. [10] investigated the task of teaching English measurement adjectives such as big and high to Turkish-speaking 5-year-olds to investigate how scaffolding (attention cues or gestures on a tablet) supports robot instruction and whether the type of gesture affects the effectiveness of the instruction The type of gesture affects the effectiveness of instruction. An interesting aspect of their study is that they also conducted experiments with human instructors to generalize their results. The results showed that children learned more vocabulary with both types of instructors and performed better when scaffolding was present. An exploratory comparison of robot and human tutors found that children in the robot tutor condition performed better than children in the human tutor condition. The authors refrain from highlighting trends indicating that the robot facilitated better learning outcomes than the human tutor, but acknowledge the usefulness of the robot tutor in second language learning.

From the above review of related studies, it is clear that attempts to directly compare RALL systems and human tutors are inadequate. This study will work to fill that gap.

3 Methods

We conducted an experiment to compare the learning outcomes of robot and human tutors in L2 learning for adults. All participants were university students. The experiment lasted for seven days, with a learning time of 30 min per day. The lessons of each day were designed to last approximately 30 min, and the participants always completed the lessons for that day. The experimental design was a two-factor mixed design in which the tutor factor (between participants) was compared before and after the 7-day lesson (within participants).

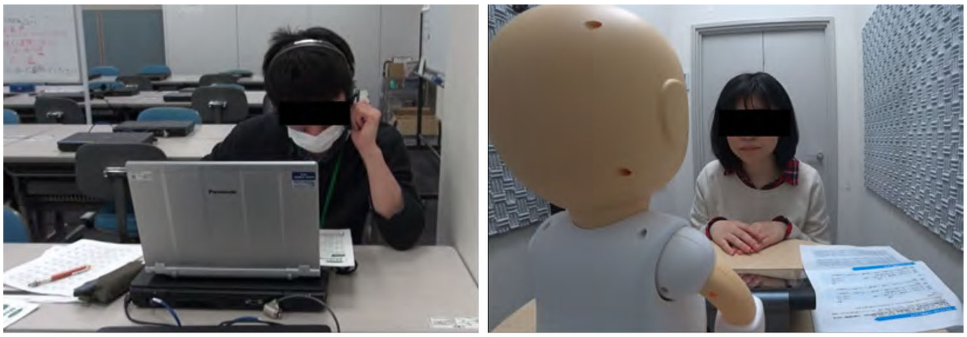

(a) Human tutoring group **(b) Robot tutoring group**

Fig. 1 Conditions of the experiment

3.1 Experimental Design

We employed a two-factor mixed design in which the tutor factor was compared before and after lessons (i.e., pre-post factor). The tutor factors comprised a RALL system and human tutor group. This was a between-participants factor, and the participants were assigned to one group or another. The pre-post factor consisted of a pre-test conducted on the first day of the experiment and a post-test conducted on the last day. This was a within-participant factor; the participants took both tests.

Figure 1 illustrates the experimental conditions. The photos on the left and right show the human and RALL system groups, respectively. Participants in the human–tutor group had online lessons with a human tutor.

3.2 Participant

A total of 26 university students (12 males and 14 females) participated in the experiment. We did not ask their age; the average age was assumed to be between 18 and 22 years considering the average age of Japanese university students. These students were recruited from the company Benesse we were collaborating with and had part-time jobs at the company. The students were recruited because it was difficult to recruit people who can participate in an experiment for seven consecutive days. These students were not involved in any work related to this study and had no knowledge of the experiment. Thus, they were not in a Conflict of interest situation. We explained the purpose and procedure of this experiment to them and obtained their informed consent. No participants refused to participate in the study.

The participants underwent a speaking test at GTEC[1] to assess their English speaking skills. The purpose was balancing the speaking ability of the robot and human tutor groups. The mean score of the participants was 115.6 (*SD* = 23.24, min = 68, max = 167). GTEC is an online test offered by Benesse Corporation[2] for reading, writing, listening, and speaking English. According to the official website, the lowest score (68) was for beginners who had difficulty speaking English even for simple things. The average score (115) was for those who could speak English in routine situations such as making a phone appointment or shopping. The highest score (167) was for those who could use English in general work, such as management, discussions, and negotiations.

The participants were assigned to the two groups such that an equal gender ratio and average GTEC score could be maintained, subject to the constraints of the participants' availability.

- *Human tutor group* The number of participants was 12 (6 males and 6 females), and the mean GTEC score was 111 (*SD* = 20.46, min = 78, max = 151).
- *RALL system group* The number of participants was 14 (6 males and 8 females), and the mean GTEC score was 116.8 (*SD* = 25.14, min = 68, max = 167).

In this experiment, we did not do power analysis because this study was exploratory. Since it has been difficult to predict the theory-based effect size and to define a specific hypothesis regarding the difference between the outcomes brought about by human and RALL systems, we considered it meaningless to design a sample size when the effect size was largely unpredictable. Therefore, we set a sample size that did not deviate from the customary range and within the constraints of human resources.

3.3 Learning Materials

Table 1 shows typical differences between the RALL system group and the human teacher group regarding lesson materials. The details of each group's materials, including the contents of this table, are described below.

[1] https://www.benesse.co.jp/gtec/en/.

[2] https://www.benesse-hd.co.jp/en/.

Table 1 Summary of the difference between the RALL system and human tutor group

	RALL system group	Human tutor group
Appearance	Humanoid robot Tablet	Human
Voice	Human (Recorded voice)	Human
Role	Lesson progression (Tablet) Role-play partner (Robot)	Lesson progression Role-play partner Instruction in learned phrases Confirmation of understanding Free talk
Environment	Face-to-face	Online (with video and audio)
Learning materials	Tablet	Paper handout
Instructions	Displayed on the tablet	Povided by human tutors
Exercises	Role-play with positive tones Reading out key expressions Role-play with negative tones Role-play **Each practice was done at least once in one lesson.*	Role-play with two answer choices Flashcards practice Role-play **Each practice was repeated four times in one lesson*

3.3.1 For the RALL System Group

We used the RALL system developed in our previous study [18]. The system was designed to improve adults' L2 speaking skills and provide three exercises based on Levelt's language processing model [30]: role play with two answer choices, flashcard practice, and role play. We provide an overview of the system and the exercises to help us understand the experiment. The details of the system and exercises were explained in our previous paper [18].

System The system consists of a tablet and a robot, as shown in Fig. 2.

The robot acts as a conversation partner during role play. The robot is a desktop humanoid robot called CommU (VSTONE), [3] and has 14 degrees of freedom (three axes for the head, three for the eyes, one for the eyelids, one for the mouth, two for the right shoulder, two for the left shoulder, and two for the waist). Human-like behaviors were implemented to enhance the sense of interactivity of students when role playing with the robot. For example, the robot nodded to indicate affirmation or backchannel, slumped to indicate consideration, and raised both hands to indicate joy or surprise. In this experiment, the robot did not use text-to-speech (TTS) technologies but played an audio file containing the speech of a native speaker because the quality of TTS was imperfect for English conversation lessons. The robot was linked to a tablet and operated in response to events.

The tablet functioned as an input–output interface for the students. The tablet received the students' button-tap events, displayed learning instructions, and played audio files. Exploiting tablets in RALL is reasonable given the current HRI technology. For example, it is difficult to control robots autonomously and socially in a complex and dynamic environment [50]. Further, it is difficult to accurately recognize speech in children [24]; moreover, it is difficult to accurately recognize speech in a second language, even in adults [11]. The guidelines for designing social robots as second-language tutors recommend that using a tablet makes the design context more flexible and reduces the need to rely on complex object recognition and tracking [7]. Therefore, in this study, we combined a tablet and a robot. This limits

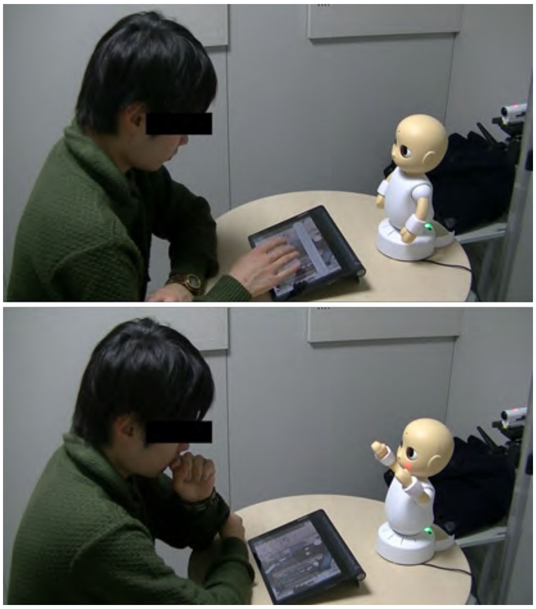
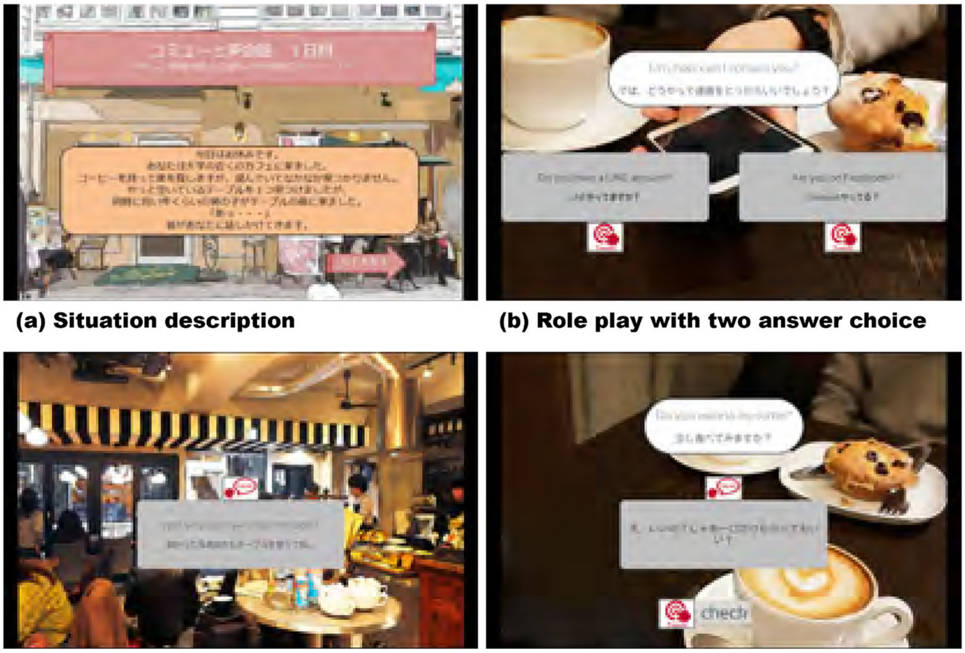

[redacted]
[redacted]

Fig. 2 RALL system. Operations on the tablet by a participant are linked to the progress of exercises

[3] https://www.vstone.co.jp/english/index.html.


(a) Situation description

(b) Role play with two answer choice

(c) Flashcards practice

(d) Role play

Fig. 3 Screenshots of exercises

the freedom of interaction and avoids the speech-recognition problem. Consequently, the curriculum can proceed formally.

Exercises Three exercises were designed based on Levelt's language processing [30].

Role Play with Two Answer Choices This exercise was designed to teach a certain conversation in a certain situation (e.g., a conversation between a man and woman seated together in a café). To achieve this objective, the robot and student played the characters according to the prepared conversation scripts. The specific learning procedure was as follows: First, the situation was presented on the tablet in Japanese (see Fig. 3a). When the Next button on the tablet was pressed, the scripts of the two characters were displayed in Japanese and English (Fig. 3b). Simultaneously, the robot spoke Character A's script in English. Character B's script was of two types: one with a positive tone and the other with a negative tone. We believe that learning both positive and negative tones would allow students to respond to more situations. When a student selected one of the scripts, the tablet played the audio of the dialogue twice. This audio was a recording of a native speaker's utterances. The audio files on the tablet and on the robot were by different speakers to enhance the realism of the role-play. Students listened to the first playback and repeated the audio during the second playback. This procedure was repeated until the conversation ended. This practice was repeated four times. As the number of sessions progressed, the level of difficulty increased, including the loss of English notation.

Flashcards Practice This exercise was designed to consolidate the scripts of Character B used in the previous memory exercise. To achieve this goal, students repeated the scripts displayed on the tablets. The specific learning procedure was as follows: First, the student chose whether to learn scripts with a positive or negative tone. From this point, only dialogues with the selected tone were studied. The scripts were displayed on the tablet in Japanese and English (Fig. 3c). When the students pressed a button on the tablet, they played the dialogue twice. The voice was the same as that used in the previous exercises. Students listened to the first playback and repeated the audio during the second playback. This procedure was repeated for all scripts. Flashcards practice was also repeated four times.

Role Play This exercise was designed to allow students to use scripts memorized through flashcard practice in conversations. To achieve this goal, the robot and student played the same characters in the same situation as before, just as in the first exercise: role play with two answer choices. However, unlike the first exercise, the English text of the script of character B was not displayed on the tablet. In other words, the students had to recall and express what they had learned. The specific learning procedure was as follows: The first script of character A was displayed on the tablet in Japanese and English, and the next script of character B was displayed only in Japanese (see Fig. 3d). The script of character B was displayed in either a positive or negative tone, whichever had been learned in flashcard practice. The robot utters the script of character A in English, and then the student utters the script of character B in English, relying on his/her memory. When the students tapped the check button on the tablet, the voice learned during the flashcard exercise was repeated twice. The student checked the first playback to see if his or her utterance was correct and then repeated the voice on the second playback. This procedure is repeated until the end of the script. This role play was also repeated four times in the same manner as the previous practices.

3.3.2 For the Human Tutor Group

Tutors' skills People are actively employed as human tutors in English conversation schools. However, it was difficult to control for the tutors' level of experience owing to the small number of tutors that we could employ for this study. This will not pose a problem as previous studies have already confirmed that tutor experience has little effect on student learning outcomes [44].

We ensured that there were no differences in the teaching styles among tutors. We worked with the coordinator of the tutors to determine lesson contents and teaching methods based on handouts and provided a forum for sharing these with other tutors. To ensure that individual tutors could perform their best, each tutor thoroughly discussed the teaching style and flow of the lesson and determined its structure. According to the English conversation school with which we collaborated, this approach is used in regular online tutoring services. Furthermore, we did not give the human tutors any instructions regarding gestures. Therefore, the human tutors were able to teach as they always do in their usual online English conversation lessons. In other words, the human tutors' performance was not limited by the experiment.

Participants did not have the same tutor for every lesson because it was extremely difficult to coordinate the schedules of participants with those of their tutors. Even if the tutor changed during the course of the lesson, the lesson procedure remained consistent because of the aforementioned information sharing.

Handout We carefully designed handouts used by tutors and students to align learning between robot and human tutors as closely as possible. The handout describes interactions and lesson content that can occur during lessons with the robot tutor. The left page of the handout contained (1) positive-tone scripts from the role play and (2) target expressions used. The right page contained (3) practice to create a sentence that included the target expressions and (4) negative-tone scripts from the role play. Thus, the handout played a similar role in the interaction as the tablet did with the robot tutor. Figure 4 shows the handouts on the first day.

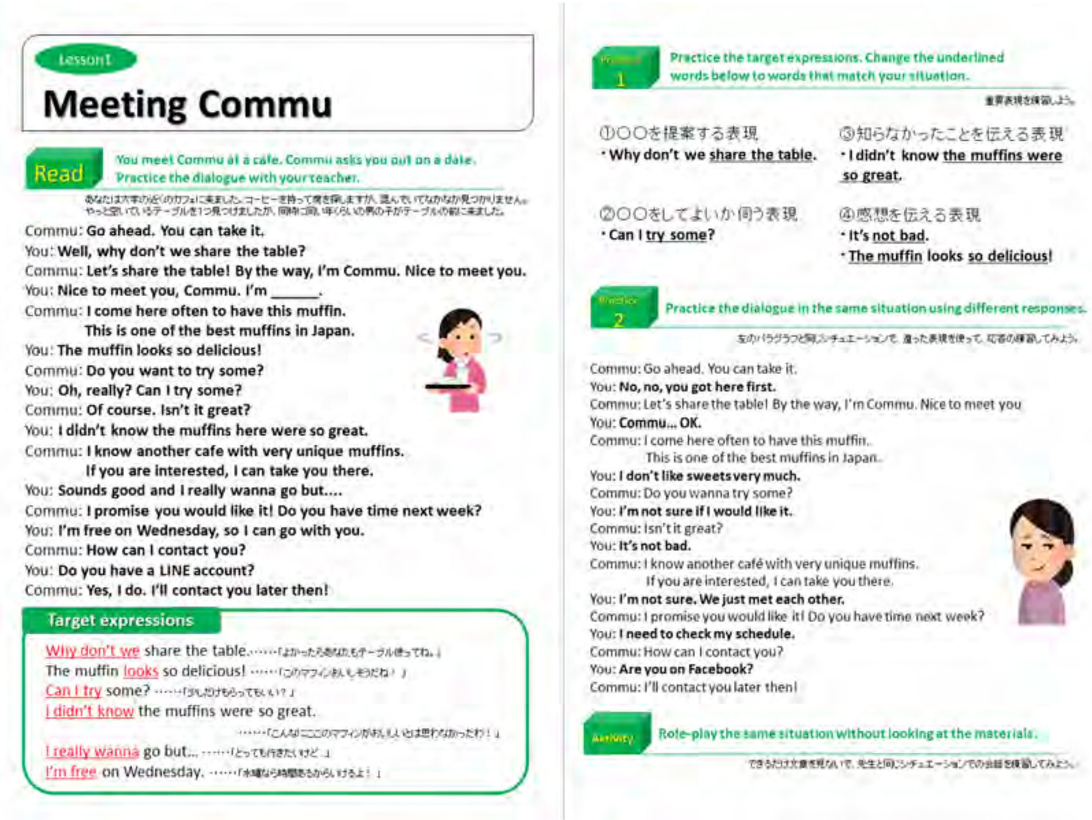
Exercises Human tutors conducted exercises that corresponded to the three exercises of the RALL system (role play with two answer choices, flashcard practice, and role play). The three types of exercises were reading scripts, reading key expressions, and role play without using text. These exercises constituted a single class, and the time allocation was not strictly controlled. The tutors could fine tune the exercises to suit the students; this was done to respect the tutors' teaching styles and maximize their performance during the lessons.

Reading Out Scripts (Positive Tones) Reading scripts is an exercise in which students learn about a particular conversation in a particular situation. Scripts on the left-hand side of the handouts were used. First, the students read the scripts silently. Next, the student played the role of “You” and the tutor played the role of “CommU” and read the scripts aloud in a role play format. Then, they changed roles and re-read the scripts. The tutor then asked simple questions to check the students' understanding of word meanings, pronunciation, and dialogue. At the tutor's discretion, explanations, exercises, or chats about the dialogue were added depending on the student's level of understanding.

Reading Out Key Expressions Reading out key expressions is an exercise for studying important expressions used in scripts and their applications. This exercise uses the important expressions listed at the bottom of the left page of the handout and their respective practices to create a sentence listed at the top of the right page. First, the student reads aloud important expressions on the bottom of the left page according to the tutor's instructions. Next, the important expressions are read aloud again, with the usages listed at the top of the right page. Then, according to the tutor's instructions, the underlined content is changed and the students practice speaking different sentences using the same expression. This exercise corresponds to the robot tutor's flashcard practice in the sense that important expressions are practiced repeatedly to be consolidated into memory.

Reading Out Scripts (Negative Tones) After reading the key expressions, the scripts on the right-hand side of the handout were read. The difference between the first and second reading was the attitude of the “You” actor's response. In the first reading out scripts, “You” responded positively to “CommU” utterances, whereas in this exercise, “You” responded negatively. The learning procedure was the same as that used during the first reading of the scripts. Through the two read-out script exercises, students learned both positive and negative responses to the speakers' suggestions. In other words, these exercises corresponded to the robot tutor's role play with two answer choices.

Role Play Role play is an exercise in which students practice dialogue in practical situations. The student role played with the tutor multiple times without looking at the handout. In this case, the tutor played the role of “CommU” and the student played the role of “You”. In the first session, the students practiced dialogue sentences on the left page, and in the second session, they practiced dialogue sentences on the right page. The number and order of the roleplays were



Lesson1

Meeting Commu

Read You meet Commu at a cafe. Commu asks you out on a date. Practice the dialogue with your teacher.

あなたは大学の近くのカフェに来ました。コーヒーを持って席を探しますが、混んでいてなかなか見つかりません。やっと空いているテーブルを1つ見つけましたが、同時に同い年くらいの男の子がテーブルの前に来ました。

Commu: **Go ahead. You can take it.**
You: **Well, why don't we share the table?**
Commu: **Let's share the table! By the way, I'm Commu. Nice to meet you.**
You: **Nice to meet you, Commu. I'm ______.**
Commu: **I come here often to have this muffin.**
This is one of the best muffins in Japan.
You: **The muffin looks so delicious!**
Commu: **Do you want to try some?**
You: **Oh, really? Can I try some?**
Commu: **Of course. Isn't it great?**
You: **I didn't know the muffins here were so great.**
Commu: **I know another cafe with very unique muffins.**
If you are interested, I can take you there.
You: **Sounds good and I really wanna go but….**
Commu: **I promise you would like it! Do you have time next week?**
You: **I'm free on Wednesday, so I can go with you.**
Commu: **How can I contact you?**
You: **Do you have a LINE account?**
Commu: **Yes, I do. I'll contact you later then!**

Target expressions

Why don't we share the table.……「よかったらあなたもテーブル使ってね。」
The muffin looks so delicious! ……「このマフィンおいしそうだね！」
Can I try some? ……「少しだけもらってもいい？」
I didn't know the muffins were so great.
……「こんなにここのマフィンがおいしいとは思わなかったわ！」
I really wanna go but… ……「とっても行きたいけど」
I'm free on Wednesday. ……「水曜なら時間あるから行けるよ！」

Practice 1 Practice the target expressions. Change the underlined words below to words that match your situation.

重要表現を練習しよう。

①〇〇を提案する表現
・**Why don't we share the table.**

②〇〇をしてよいか伺う表現
・**Can I try some?**

③知らなかったことを伝える表現
・**I didn't know the muffins were so great.**

④感想を伝える表現
・**It's not bad.**
・**The muffin looks so delicious!**

Practice 2 Practice the dialogue in the same situation using different responses.

左のパラグラフと同じシチュエーションで、違った表現を使って、応答の練習してみよう。

Commu: Go ahead. You can take it.
You: **No, no, you got here first.**
Commu: Let's share the table! By the way, I'm Commu. Nice to meet you
You: **Commu… OK.**
Commu: I come here often to have this muffin.
This is one of the best muffins in Japan.
You: **I don't like sweets very much.**
Commu: Do you wanna try some?
You: **I'm not sure if I would like it.**
Commu: Isn't it great?
You: **It's not bad.**
Commu: I know another café with very unique muffins.
If you are interested, I can take you there.
You: **I'm not sure. We just met each other.**
Commu: I promise you would like it! Do you have time next week?
You: **I need to check my schedule.**
Commu: How can I contact you?
You: **Are you on Facebook?**
Commu: I'll contact you later then!

Activity Role-play the same situation without looking at the materials.

できるだけ文章を見ないで、先生と同じシチュエーションでの会話を練習してみよう。

Fig. 4 Sample of handouts of exercises

left to the tutor's discretion. This exercise corresponded to the role play of the robot tutor.

3.4 Procedure

An overview of the experimental procedure is presented in Fig. 5. Participants arrived at the laboratory for 7 d. On day 1, they took a pretest to measure their speaking skills. Then, they studied for 30 min in the assigned group, and they continued to study for 30 min on days 2 through 6. On day 7, they studied for 30 min and then took three different post-tests. The pre-test, post-test, and exercise were conducted in a private room or in an experimental booth built as part of a larger room. We set up video cameras in the private room and the experimental booth to record the participants' audio and behavior. The experimental procedures were approved by the Osaka University Ethics Committee (approval number: 31-2-2).

3.5 Pre-test and Post-tests

The pre- and post-tests were conducted to evaluate the participants' speaking skills in conversations before and after the study. The participants' responses to their partners' utterances in certain conversational situations were evaluated. The pre- and post-tests had the same format, and both were created using Microsoft PowerPoint®. The first slide showed instructions on how to perform the test, and the test began when the participant advanced to the next slide. The second slide displayed the conversational situation, and when the slide advanced to the third slide, the voice file was played. After the voice was played, the participant was signaled to begin responding to the utterance. The participants repeated this procedure 15 times (i.e., 15 utterances were evaluated). All participants' utterances were recorded.

The participants took three different post-tests on the last day. Post-test A was identical to the pre-test. Post-test B had the same conversational situation as Post-test A, but with a different vocabulary. Post-test C differed from Post-test A in terms of both conversational context and vocabulary. Post-test A was identical to the pre-test and was conducted to assess improvement in speaking skills before and after the study. Post-tests B and C were conducted to assess whether the participants were able to apply what they had learned.

3.6 Evaluation

The recordings of the participants' speeches on the pre-test and post-tests were evaluated by an expert in English language education for Japanese students from the company

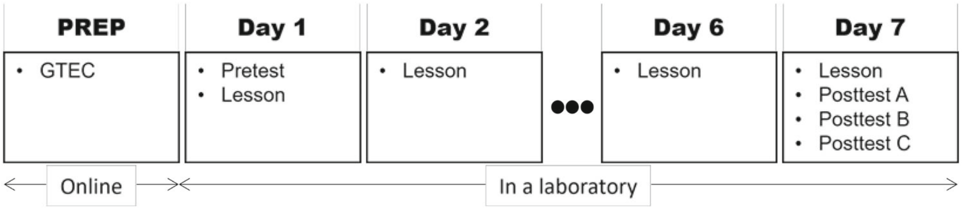


Fig. 5 Experimental procedure

with which we collaborated. To reduce the burden of evaluation, the evaluation was shared among several experts. Because each participant was evaluated by any one expert, we could not calculate the consistency of the ratings among the experts. However, since the experts are current employees of online English schools, we believe that the consistency of their ratings is reliable. The evaluation is done for each question response in those tests. The final score for a participant is the normalized sum of the scores for each response. Specifically, if the number of questions is 15 and the rating is on a scale of 3 (0,1 or 2), the participant's total score will be a value ranging from 0 to 30, and the final score will be that value normalized from 0 to 1. Arguments regarding the use of evaluation were discussed in our previous work [18]. Here, we explain the methods used for evaluation. The evaluation is based on the following six perspectives.

3.6.1 Error

Error was evaluated using lexical and grammatical error rates. The error rates were calculated as W_m/W_s, where W_m and W_s denote the numbers of words missed and uttered, respectively. For example, the sentence "This is pen" contains a grammatical error because the "a" is missing. The error rate for this sentence is 1/3 = 33%. A lower error rate indicates a more accurate utterance.

3.6.2 Fluency

Fluency was evaluated based on the number of words uttered per second. The number is calculated as W_s/D_s, where D_s denotes the utterance duration.

3.6.3 Rhythm

Rhythms were evaluated using a third-person subjective assessment. Native English speakers rated the rhythm of each participant's speech on a three-point scale. Natural speech rhythm was rated as "good", slightly unnatural but understandable speech rhythm as "so-so", and incomprehensible speech rhythm as "not good".

3.6.4 Pronunciation

Pronunciations were evaluated using a third-person subjective assessment. Native English speakers rated the quality of the pronunciation of words of each participants' speech on three levels: "native", "some problems", and "non-native".

3.6.5 Complexity

Complexity was evaluated based on the number of words per unit of AS (Analysis of Speech [12]). The number was calculated as W_s/A_s, where A_s represents the number of AS units in speech. The AS unit is a single utterance consisting of independent phrases [12]. This is almost synonymous with the sentences. For example, "I play tennis" is an AS unit consisting of three words. The complexity of this sentence is 3/1 = 3. A higher complexity score means that a participant can speak longer sentences without pauses.

3.6.6 Task Achievement

Task achievement was evaluated using a third-person subjective assessment. Native English speakers rated whether each participants' response was appropriate to a question or suggestion on a two-point scale of "appropriate" or "inappropriate".

3.7 Analysis

A two-factor mixed-design analysis of variance was used to compare the pre-test results with post-test A results across all tutor factors. If the test indicated an interaction, Tukey's post-hoc test was conducted to test the difference between each level.

Welch's t-test was used to compare tutor factors between the robot and human tutors in post-tests B and C.

The alpha level was set at 0.05 in both tests.

Table 2 Descriptive statistic of participants' scores in the pre-test and post-test A

Measures	Range	Test	Condition	N	Mean	95% CI		SD
						Lower	Upper	
Error	[0,1]	Pre	Human	12	0.094	0.0593	0.128	0.054
			Robot	14	0.103	0.065	0.141	0.066
		Post	Human	12	0.055	0.0262	0.084	0.045
			Robot	14	0.011	0.001	0.020	0.016
Fluency	(0,∞)	Pre	Human	12	1.489	1.108	1.870	0.600
			Robot	14	1.469	1.168	1.769	0.521
		Post	Human	12	1.702	1.381	2.022	0.505
			Robot	14	2.247	1.985	2.509	0.454
Rhythm	[0,1]	Pre	Human	12	0.922	0.844	1.001	0.123
			Robot	14	0.691	0.592	0.789	0.171
		Post	Human	12	0.933	0.889	0.978	0.070
			Robot	14	0.967	0.930	1.003	0.063
Pronunciation	[0,1]	Pre	Human	12	0.589	0.476	0.701	0.177
			Robot	14	0.391	0.247	0.534	0.248
		Post	Human	12	0.633	0.440	0.827	0.304
			Robot	14	0.520	0.385	0.653	0.233
Complexity	[1,∞)	Pre	Human	12	4.163	3.617	4.710	0.860
			Robot	14	4.155	3.562	4.749	1.028
		Post	Human	12	4.535	4.095	4.975	0.693
			Robot	14	4.669	4.333	5.005	0.581
Task achieve	[0,1]	Pre	Human	12	0.939	0.897	0.981	0.066
			Robot	14	0.943	0.901	0.985	0.073
		Post	Human	12	0.967	0.938	0.995	0.045
			Robot	14	0.981	0.957	1.005	0.041

Range means the range of possible values of each measures. Fluency and Complexity have biological limits but no upper limit in their definition

4 Results

4.1 Pre-test and Post-test A

Table 2 shows the descriptive statistics of the participants' scores in the pre-test and post-test A, and Fig. 6 plots the scores for each participant.

4.1.1 Error

The two-factor mixed-design ANOVA showed a main effect of the pre-post factor ($F(1, 24) = 35.15, p < .001, \eta^2 = 0.302$), no main effect of the tutor factor ($F(1, 24) = 1.26, p = .273, \eta^2 = 0.022$), and an interaction between these factors ($F(1, 24) = 5.89, p < .023, \eta^2 = 0.051$). Tukey's post-hoc test showed the following results:

- The mean of the error rates of post-test A was significantly lower than that of the pre-test in the RALL system group ($t = 6.150$, $p < .001$).
- No significant difference was found between the pre-test and post-test A in the human tutor group ($t = 2.385$, $p = .107$).
- No significant difference was found between the human and RALL system groups in the pre-test ($t = -0.382$, $p = .981$).
- The mean of the error rates of the RALL system groups was significantly lower than that of the human tutor groups in post-test A ($t = 3.439$, $p = .011$).

4.1.2 Fluency

The two-factor mixed-design ANOVA showed a main effect of the pre-post factor ($F(1, 24) = 24.12, p < .001, \eta^2 = 0.176$) and an interaction ($F(1, 24) = 7.85, p = .010, \eta^2 = 0.057$), but no main effect of the tutor factor ($F(1, 24) = 2.18, p = .153, \eta^2 = 0.049$). Tukey's post-hoc test showed the following results.

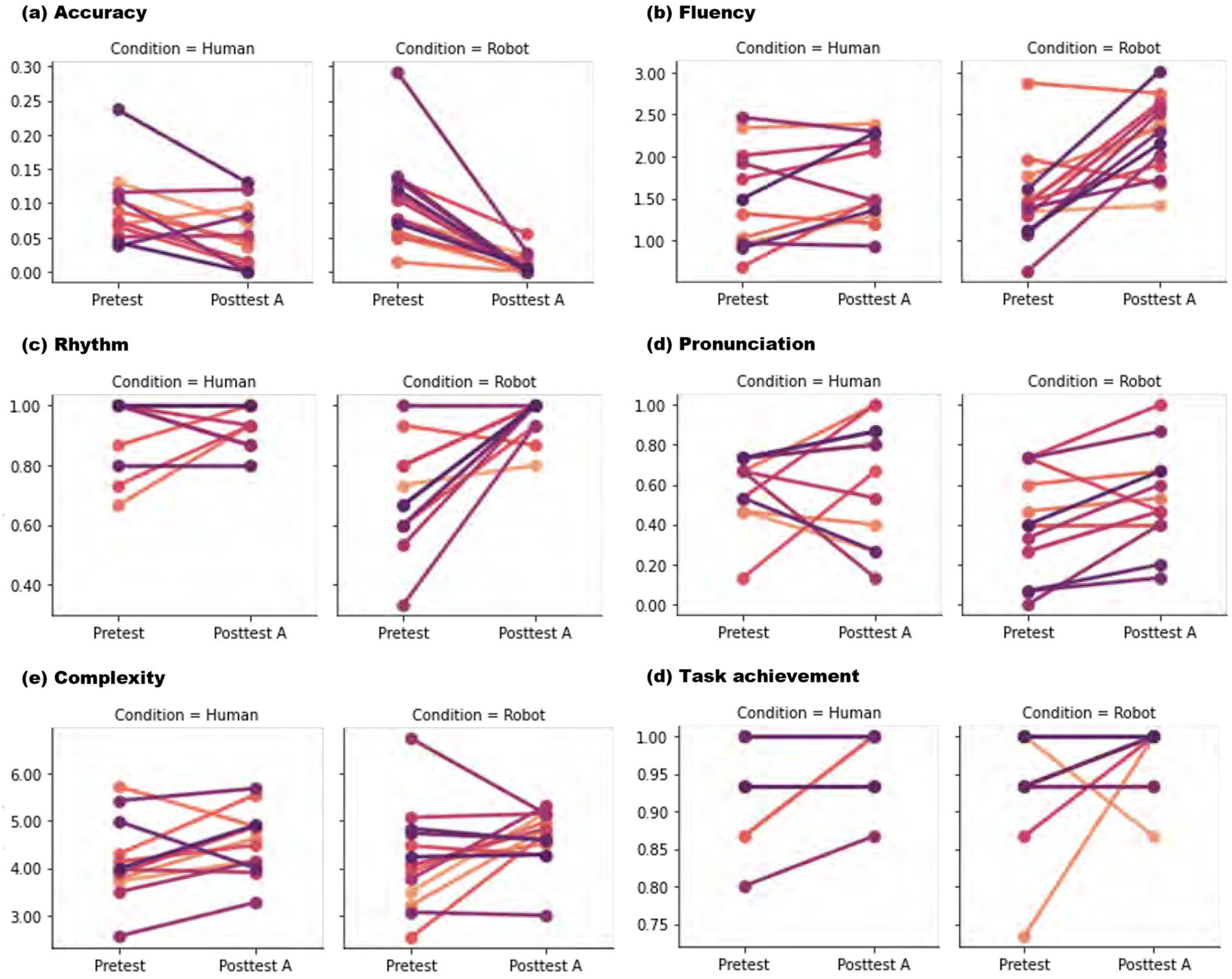


Fig. 6 Plots of all participants' scores in the pre-test and post-test A

- The mean of the number of words uttered per second of post-test A was significantly higher than that of the pre-test in the RALL system group ($t = -5.677$, $p < .001$).
- No significant differences were found between the pre-test and post-test A in the human tutor group ($t = -1.438$, $p = .489$).
- No significant difference was found between the human and RALL system groups in the pre-test ($t = 0.092$, $p = 1.000$).
- The mean of the number of words uttered per second of the RALL system group was significantly higher than that of the human tutor group in post-test A ($t = -2.899$, $p = .037$).

4.1.3 Rhythm

The two-factor mixed-design ANOVA showed a main effect of the pre-post factor ($F(1, 24) = 20.1$, $p < .001$, $\eta^2 = 0.21$), a main effect of the tutor factor ($F(1, 24) = 9.2$, $p = .006$, $\eta^2 = 0.1$), and an interaction ($F(1, 24) = 17.1$, $p < .001$, $\eta^2 = 0.179$). Tukey's post-hoc test showed the following results:

- The mean of the scores of post-test A was significantly higher than that of the pre-test in the RALL system group ($t = -6.342$, $p < .001$).
- No significant differences were found between the pre-test and post-test A in the human tutor group ($t = -0.236$, $p = .995$).
- The mean of the scores of the human tutor group was significantly higher than the RALL system group in the pretest ($t = 3.905$, $p = .003$).
- No significant difference was found between the human and RALL system groups in post-test A ($t = -1.284$, $p = .581$).

4.1.4 Pronunciation

The two-factor mixed-design ANOVA showed no main effect of the pre-post factor ($F(1, 24) = 3.331$, $p = .080$, $\eta^2 = 0.029$), no main effect of the tutor factor ($F(1, 24) = 3.50$, $p = .074$, $\eta^2 = 0.096$), and no interaction ($F(1, 24) = 0.787$, $p = .384$, $\eta^2 = 0.007$).

4.1.5 Complexity

The two-factor mixed-design ANOVA showed a main effect of the pre-post factor ($F(1, 24) = 6.454$, $p = .018$, $\eta^2 = 0.074$), no main effect of the tutor factor ($F(1, 24) = 0.055$, $p = .816$, $\eta^2 = 0.001$), and no interaction ($F(1, 24) = 0.165$, $p = .688$, $\eta^2 = 0.002$).

4.1.6 Task Achievement

The two-factor mixed-design ANOVA showed a main effect of the pre-post factor ($F(1, 24) = 4.989$, $p = .035$, $\eta^2 = 0.079$), and no main effect of the tutor factor ($F(1, 24) = 0.273$, $p = .606$, $\eta^2 = 0.006$) or interaction ($F(1, 24) = 0.122$, $p = .729$, $\eta^2 = 0.002$).

4.2 Post-tests B and C

Table 3 shows the descriptive statistics of the participants' scores in post-tests B and C. Figure 7 plots the scores for each participant.

4.2.1 Error

Welch's t-test showed that the mean error rates of the RALL system group were significantly lower than those of the human tutor group in post-tests B ($t(24.0) = 2.317$, $p = .029$, $Cohen'sd = 0.905$) and C ($t(15.7) = 3.109$, $p = .007$, $Cohen'sd = 1.251$).

4.2.2 Fluency

Welch's t-test showed no significant differences between the robot and human tutor groups in post-tests B ($t(22.9) = -1.834$, $p = .080$, $Cohen'sd = -0.723$) and C ($t(23.5) = -1.990$, $p = .058$, $Cohen'sd = -0.783$).

4.2.3 Rhythm

Welch's t-test showed no significant differences between the robot and human tutor groups in post-tests B ($t(13.3) = -1.337$, $p = .203$, $Cohen'sd = -0.542$) and C ($t(17.5) = -1.698$, $p = .107$, $d = -0.680$).

4.2.4 Pronunciation

Welch's t-test showed no significant differences between the robot and human tutor groups in post-tests B ($t(24.0) = -0.487$, $p = .631$, $Cohen'sd = -0.191$) and C ($t(23.9) = 0.725$, $p = .475$, $d = 0.284$).

4.2.5 Complexity

Welch's t-test showed no significant difference between the robot and human tutor groups in post-tests B ($t(16.6) = 0.690$, $p = .500$, $Cohen'sd = 0.277$) and C ($t(21.1) = 1.795$, $p = .087$, $Cohen'sd = 0.712$).

4.2.6 Task Achievement

Welch's t-test showed no significant difference between the robot and human tutor groups in post-tests B ($t(14.0) = -1.678$, $p = .116$, $Cohen'sd = -0.678$) and C ($t(22.6) = -0.674$, $p = .507$, $Cohen'sd = -0.266$).

5 Discussion

5.1 Implication

5.1.1 Error

The results indicate that learning with a RALL system can colorredreduce speaking errors more than that when learning with a human tutor. First, the lexical and grammatical error rates were not significantly different between the human and RALL system groups on the pre-test. The graph (Fig. 6a) also shows that the error rates in the pre-test of these groups are the same. Second, for both groups, the means were significantly lower on post-test A than on pre-test A, which indicates that both groups had reduced lexical/grammatical errors. However, in post-test A, the mean error rate of the RALL system group was significantly lower than that of the human tutor group. This indicates that the RALL system group reduced their errors more than that of the human tutor group. Further, the graph shows a greater decrease for the RALL system group than for the human tutor group.

The mean error rates of the RALL system group were significantly lower than those of the human tutor group in post-tests B and C. The effect sizes were 0.897 and 1.422 for post-tests B and C, respectively. We believe that the effect sizes were generally large; these results indicate that learning with a RALL system reduces errors of utterances more than that with human tutors in advanced tests.

Table 3 Results of speaking skills

Measures	Range	Test	Condition	N	Mean	95% CI		SD
						Lower	Upper	
Error	[0,1]	Posttest B	Human	12	0.050	0.036	0.064	0.022
			Robot	14	0.028	0.012	0.043	0.027
		Post-test C	Human	12	0.043	0.024	0.062	0.029
			Robot	14	0.014	0.006	0.023	0.015
Fluency	(0,∞)	Post-test B	Human	12	1.638	1.320	1.957	0.502
			Robot	14	1.992	1.717	2.267	0.476
		Post-test C	Human	12	1.547	1.234	1.860	0.492
			Robot	14	1.934	1.647	2.221	0.496
Rhythm	[0,1]	Post-test B	Human	12	0.9	0.811	0.989	0.141
			Robot	14	0.957	0.929	0.986	0.050
		Post-test C	Human	12	0.894	0.823	0.966	0.112
			Robot	14	0.957	0.918	0.996	0.067
Pronunciation	[0,1]	Post-test B	Human	12	0.578	0.448	0.707	0.204
			Robot	14	0.619	0.487	0.751	0.229
		Post-test C	Human	12	0.7	0.549	0.851	0.237
			Robot	14	0.629	0.476	0.782	0.265
Complexity	[1,∞)	Post-test B	Human	12	4.544	4.006	5.082	0.847
			Robot	14	4.354	4.081	4.627	0.473
		Post-test C	Human	12	4.676	4.116	5.236	0.881
			Robot	14	4.107	3.697	4.516	0.710
Task achieve	[0,1]	Post-test B	Human	12	0.972	0.944	1.001	0.045
			Robot	14	0.995	0.985	1.006	0.018
		Post-test C	Human	12	0.972	0.950	0.994	0.034
			Robot	14	0.981	0.963	0.999	0.031

5.1.2 Fluency

For fluency, we reached the same conclusion as for error. The results indicated that learning with a RALL system may improve speech fluency compared to learning with a human tutor. In the pre-test, there was no significant difference in the number of words uttered per second between the human and RALL system groups. The graph shows that the human tutor group was slightly more scattered than the RALL system group; however, the averages were almost identical. In both groups, the number of words uttered per second was significantly higher in post-test A than in the pre-test. Speech fluency improved in both tutor groups. However, in post-test A, the number of words uttered per second of the RALL system group was significantly higher than that of the human tutor group. This suggests that the RALL system group improved their fluency more than the human tutor group. Further, the graph shows that the RALL system group exhibited a greater increase in fluency than the human tutor group.

In post-tests B and C, there was no significant difference in the number of words uttered per second between the robot and human tutor groups. Thus, it cannot be concluded if learning with a RALL system improved speech fluency more than that when learning with a human tutor, even in advanced tests. However, the graph shows that the RALL system group tended to speak faster than the human tutor group in both post-tests B and C. We believe that these data suggest that RALL systems can improve speech fluency more than human tutors in applied situations.

5.1.3 Rhythm

There was no clear evidence of a difference in the outcomes between human and RALL systems regarding speaking rhythm. The results showed an interaction between the tutor factor and the pre-post factor; however, there were no significant differences between the groups in post-test A. Further, the graph shows that both groups exhibited similar levels of rhythmic goodness. The RALL system group showed a significant difference between pre-test and post-test A, whereas the human tutor group did not. However, these results should not be interpreted to indicate a larger effect on the RALL system group. This is because, at the time of the pre-test, the mean good rate of the RALL system group was significantly lower than that of the human tutor group. It is appropriate

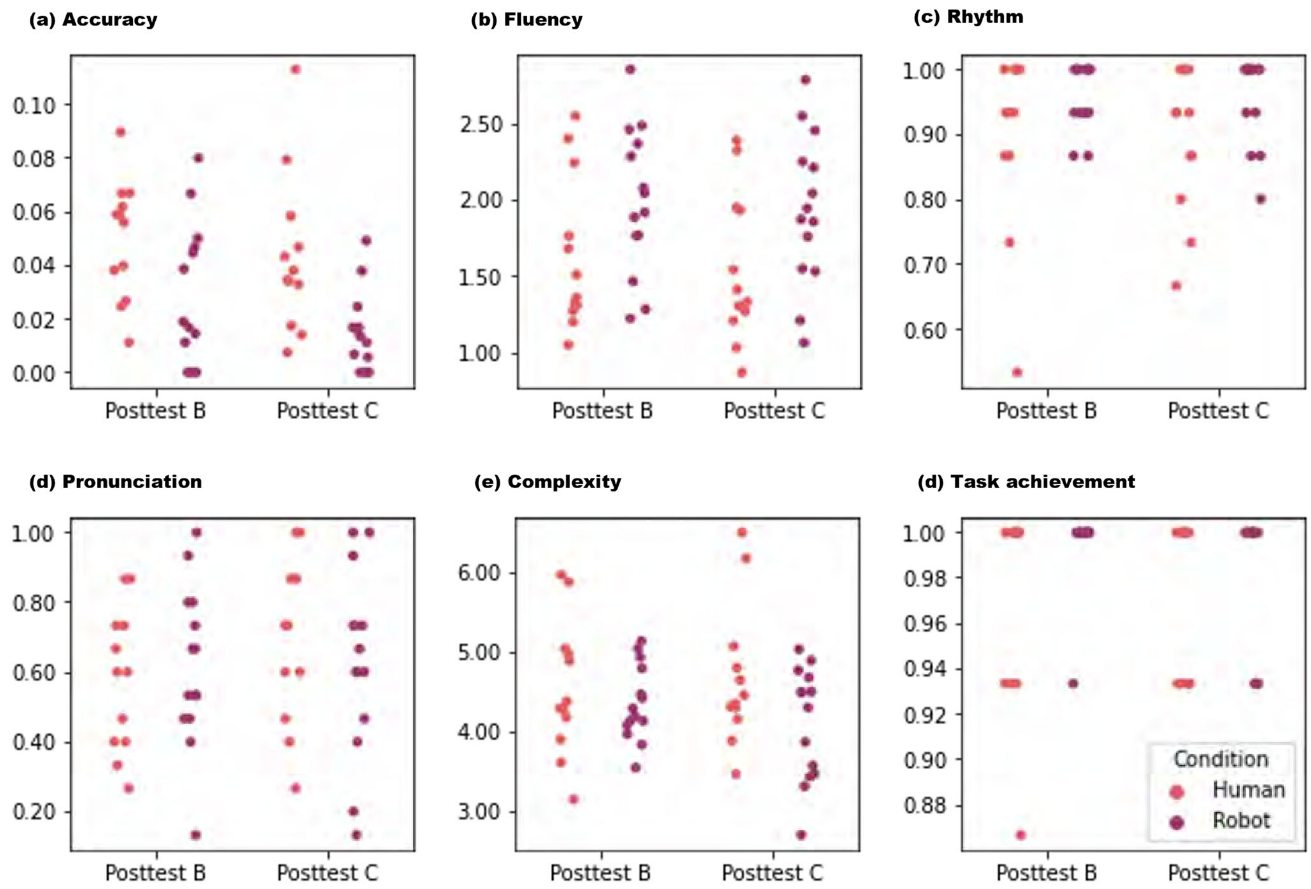


Fig. 7 Results of each measurement in post-tests B and C. The points indicate the data of each participant

to interpret this as a greater growth potential for the RALL system group. In light of the above, it is difficult to conclude that there is a difference in outcomes between the human and RALL system groups.

However, some results suggest that RALL systems are likely to lead to higher outcomes. Although the results of post-tests B and C showed no significant differences between the human and RALL system groups, the graph shows consistent data that the RALL system group had a higher mean and smaller variance than the human tutor group. Considering that the RALL system group had a lower mean in the pre-test, this suggests that the RALL system group may have achieved higher learning outcomes than those of the human tutor group.

5.1.4 Pronunciation, Complexity, and Task Achievement

No differences were found in the outcomes of the human and RALL system groups regarding pronunciation, complexity, or task achievement. Therefore, we discuss the other characteristics observed in the graphs.

Pronunciation Some characteristics suggest that the RALL system group may be more effective than the human tutor group. For example, the human tutor group had greater variance in post-test A than in the pre-test and some participants decreased their score in post-test A. However, the RALL system group showed that the variance remained the same between pre-test and post-test A and most participants improved their scores.

Complexity Several interesting features were observed. First, in post-test A, the number of words per unit of AS of the RALL system group converged around a certain point. A similar phenomenon was observed in post-test B, in which the conversational situation was the same. We believe that this phenomenon may be attributed to most of the participants who used the expressions they had learned in the post-tests. The RALL system may be more effective than human tutors in fostering a steady output of learned expressions in the test. However, for some participants who produced somewhat complex sentences, RALL systems may have prevented the improvement of their speaking skills in terms of complexity. For example, one participant in the RALL system group uttered complex sentences in the pre-test but converged on sentences using learned expressions in post-test A. Two participants in the human tutor group uttered sentences with greater complexity in post-tests B and C. Lessons with human tutors may include chitchat and human tutors' empirical judgment in teaching applied expressions. We speculate that these results may have led to complex utterances in post-tests B and C.

Task Achievement The scores of task achievement were improved in both groups; however, there was no effect in the tutor factor and no features that would indicate a difference between the two groups was found in the graphs. Note that the scores of task achievement were relatively high even in the pretest. To obtain stable scores for other items such as Lexical/Grammatical error and Fluency, the task difficulty must be set at a level that participants can respond to without much trouble. The fact that task achievement was high suggests that the difficulty level of the pretest and posttests were designed well.

5.1.5 Summary

The findings for each measurement item are summarized as follows.

- Error rates improved in both learning with the robot and human tutors. Learning with the RALL system improved error rates more than that with the human tutor.
- Fluency improved in both learning with the robot and human tutors. Learning with the RALL system improved fluency more than that with the human tutors.
- Rhythm improved in both learning with the robot and human tutors. We could not say their outcomes are different. However, some data suggested that learning with the RALL system may be more effective than learning with the human tutors.
- Pronunciation did not improve in both learning with the robot and human tutors. We found no differences in the outcomes between the robot and human tutors. However, learning with the RALL system tended to improve pronunciation more consistently than learning with human tutors.
- Complexity improved in both learning with the robot and human tutors. We found no differences in the outcomes between the robot and human tutors. However, some data suggested that learning with human tutors may have been more effective in utilizing more complex utterances in advanced tests.
- Task achievement improved in both learning with the robot and human tutors. We found no differences in the outcomes between the robot and human tutors. The raw data also did not reveal any notable characteristics.

5.2 Why did the RALL System Produce Better Outcomes of Error and Fluency than Human Tutors?

Exercises with the RALL system involved many repetitions of vocalizing expressions (especially shadowing). Such exercises encourage the consolidation of expressions practiced in memory. In the "Role play with two answer choices" and "Flashcards practice" exercises, participants repeated the same expressions. In addition, in the "Role play" exercise, the participants vocalized the expressions they had learned in the previous exercises from memory without looking at the sentences. We believe that basic training may have helped the participants retain the expressions in their memory and resulted in them recalling the exact expressions quickly during the post-tests.

However, learning with human tutors contained fewer repetitions of vocalization of expressions and fewer exercises to have participants commit the practiced expressions to memory than learning with the robot tutor. The repetition of vocal exercises and memorization of expressions comprise basic and individualized training. It is not cost effective to conduct such training with human tutors. Basic training can only be performed alone; however, practical communication can be achieved with only a human tutor. Therefore, learning with human tutors may have involved less basic training and more communication with participants. Although such communicative training improves error rates and fluency to some extent, it may not promote memory retention compared with the basic training provided by the robot tutor.

Furthermore, the participants' tension and social anxiety may explain the difference in effectiveness between the RALL system and the human tutors. Participants in the human tutor group may have felt that they were constantly being evaluated in some way by the human tutors. For example, participants may have felt that the human tutors thought that their grammar was messed up or their pronunciation was bad. On the other hand, the participants in the RALL system group would not have felt such tension or social anxiety. This is because the robot instructors did not change their facial expressions, tone of voice, or other behaviors in response to the participants' speech in any way. As a result, participants in the RALL system group were able to focus on speaking English, which may have contributed to lower vocabulary errors and increased fluency.

Based on the above discussion, we believe the RALL system may have been able to make the participants remember the learned expressions better than the human tutor.

5.3 Did the Better Error Rates and Fluency Outcomes Occur Because of RALL Systems?

The differences in the results cannot be simply reduced to differences in attributes inherent to the tutors, such as appearance and voice quality, but should be reduced to differences in their overall nature, including aspects of competence, such as what exercises they were able to provide.

It would be difficult for human tutors to perform the same exercises as RALL systems. The repeated practice of memorizing expressions is a boring exercise for both students and human tutors. Students may feel that it is a waste to assign a (costly) tutor to something they can do alone, and they and may feel uncomfortable about making intelligent human tutors go through boring exercises. Further, human tutors also want to engage in communicative exercises because they are proud of their interactive tutoring skills. Therefore, learning with human tutors did not motivate students or human tutors to perform the same exercises as that with the robot tutor.

Further, it is difficult for students to perform memory consolidation exercises alone. As discussed previously, this type of exercise requires patience. Students who train alone using smartphones or PC are tempted to stop halfway. According to the RALL studies, RALL systems can increase student compliance [6]. In other words, it is likely to reduce the urge to stop the exercise midstream. In this study, participants were asked to study in the laboratory, and we were unable to test the effect of suppressing the urge to stop the exercises. We believe that the presence of the robot may contribute to strengthening the will to continue with the exercises.

Thus, we believe that the better error rates and fluency outcomes were likely brought about by exercises that promoted memory retention, and that such exercises worked well because of the robot tutor.

5.4 Application and Limitation

The extent to which the findings of this study can be applied is discussed in terms of language type, participant demographics, learning content, robot type, and AI technology.

Language This study dealt with English. The findings of this study are likely to be applied to speaking practices in other languages such as Chinese, French, and Spanish. The retention of basic phrases in memory through repeated utterances is basic training, regardless of language type. We believe that in other languages, learning with a RALL system would be more effective for basic training than learning with human tutors.

Participant Attributes This study employed university students whose native language is Japanese. The findings of this study are probably applicable to children, middle-aged people, and elderly people other than university students. The exercises conducted in this study were simple and could be practiced easily by both children and the elderly once they become familiar with them. However, the findings of this study are not applicable to people who can create complex sentences in English. Because these people would achieve high scores even before learning, it would be difficult to find differences in the outcomes of studying with each tutor.

Learning Materials In this study, we created learning materials that emphasized role play for speaking practice. This learning material may have had a considerable impact on the present results because it maximized the advantages of the RALL system over human tutors. If the learning materials were free-talk, the results for error rates and fluency might have been different. Therefore, the findings of this study are limited to the use of learning materials that emphasize role play, including the repeated practice of basic expressions.

Robot Type This study used a table-top robot called “CommU”. Because previous studies have not reported consistent findings regarding robot appearance and learning effectiveness, it is unclear whether other types of robots would produce results similar to those in this study. We speculate that lifelike robots such as Nao, Pepper, and Tega could produce results similar to those of this study. As one of the implications of this study is the effectiveness of repeated practice through role-play, it is important that RALL systems make students feel that they are monitoring the students and can behave as partners in role play.

Audio Variation Because it was difficult to match participants’ and human tutors’ schedules, participants received lessons from more than one human tutor; according to Barcroft et al. audio variation has a positive effect on second language vocabulary learning [3]. In this regard, the participants in the human tutor group may have had a better effect than the participants in the RALL group, who only had two different voices (one for the robot and one for the tablet). In order to discuss such effects in depth, it is necessary to experiment with different types of human voices in the RALL system under controlled conditions.

Physical Presence In this experiment, physical presence could not be controlled between the RALL system and human tutor conditions. In the RALL system, the robot was in front of the participants, whereas the human tutor was online with a video display. HRI’s previous studies have shown that the physical body of a robot has positive effects in interaction. Given these findings, the difference between the human and RALL system conditions might have been smaller if the human tutor had been in front of the participants. It should be added, however, that even if such a result were obtained, it would not make the findings of this study any less meaningful. This is because the physical face-to-face learning between the human tutor and the learner is extremely costly, and the actual learning is mostly online. In this sense, the findings of this study provide useful insights into the actual situation.

AI Techniques In this study, the robot behavior is based on classical scenario-based techniques and does not use newer techniques such as large-scale language models or personalization. Even with these newer techniques, the findings of this study would still be useful. If new technologies are used, RALL systems can provide adaptive instructions that are more similar to those of human tutors. However, the RALL systems are still machines. It is likely that RALL systems will make it easier for students to request repeated practice than with human tutors.

If technology develops further and robots and humans become almost indistinguishable, students may become uncomfortable with robots. In the future, it may be necessary to change the level of humanness and intelligence perceived by students between partner robots for basic and advanced practices.

Combination of Human and RALL Systems We believe that a combination of human and RALL systems will produce the best results for English learning. However, the role of the RALL system will increase as AI and robot technology advances. At the time this study is conducted, it would be appropriate for the RALL system to provide basic training, such as repeated utterances of key phrases. This is because the technology to accurately recognize non-native speakers' speech, to synthesize speech like a native speaker, and to understand the intent of the learner's utterances is not sufficient. Instead, human tutors should conduct classes and open-ended dialogues that proceed interactively according to the learner's situation. When the above technologies are sufficiently developed, RALL systems will be able to conduct the interactive lessons and open-ended dialogues that human tutors have been doing until then. Robots will be able to replace most of the exercises in English conversation learning. However, this does not mean that human tutors will be unnecessary. If the purpose of learning English conversation is to communicate with English-speaking people, then real human communication practice will still be necessary. Some English conversation learners may feel nervous or anxious about communicating with others in a language with which they are unfamiliar. Practicing communication with real people will be indispensable to get used to such nervousness and anxiety.

6 Conclusion

This study compared learning outcomes between robot and human tutors, with a focus on speaking skills in second-language learning. Through a 7-day experiment with a two-factor (tutor and pre-post factors) mixed design, we found that participants who learned with a RALL system significantly improved their error rates and fluency in speech compared to those who learned with a human tutor. No significant differences were observed between the robot and human tutors in terms of rhythm, pronunciation, complexity, or task achievement. These results are derived from the fact that participants in the RALL system group addressed repeated practice for vocalizing expressions more than those of the human tutors. Such practices can help students retain learned expressions in their memories. This practice, which involves many repetitions, is difficult to perform with human tutors, and therefore, RALL systems that can provide it have an advantage.

Thus, we conclude that RALL systems may be more effective than human tutors in helping students memorize basic phrases in their second language speaking skills. This study demonstrated the benefits of using RALL systems that can work well in lessons that are hard to teach for human tutors.

Acknowledgements We thank Benesse Corporation for their assistance with this study. Tutors who were native English speakers from Benesse Corporation helped in the assessment of the participants' voice data.

Funding This work was supported by JST Moonshot R&D Grant Number JPMJMS2011 (Analysis), JSPS KAKENHI Grant Number 19H05691 (Development of the robot system and the tablet interface), and the Benesse Corporation (Development of the learning materials and assessment of the tests).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Alemi M, Meghdari A, Ghazisaedy M (2014) Employing humanoid robots for teaching English language in Iranian junior high-schools. Int J Humanoid Rob 11(03):1450022
2. Alemi M, Meghdari A, Ghazisaedy M (2015) The impact of social robotics on l2 learners' anxiety and attitude in English vocabulary acquisition. Int J Soc Robot 7(4):523–535
3. Barcroft J, Sommers MS (2005) Effects of acoustic variability on second language vocabulary learning. Stud Second Lang Acquis 27(3):387–414
4. Barsalou LW et al (2008) Grounded cognition. Annu Rev Psychol 59(1):617–645

5. Baxter P, Ashurst E, Read R, Kennedy J, Belpaeme T (2017) Robot education peers in a situated primary school study: personalisation promotes child learning. PLoS ONE 12(5):e0178126
6. Belpaeme T, Kennedy J, Ramachandran A, Scassellati B, Tanaka F (2018) Social robots for education: a review. Sci Robot 3(21):eaat5954
7. Belpaeme T, Vogt P, Van den Berghe R, Bergmann K, Göksun T, De Haas M, Kanero J, Kennedy J, Küntay AC, Oudgenoeg-Paz O et al (2018) Guidelines for designing social robots as second language tutors. Int J Soc Robot 10:325–341
8. Chang C-W, Lee J-H, Chao P-Y, Wang C-Y, Chen G-D (2010) Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school. J Educ Technol Soc 13(2):13–24
9. de Wit J, Vogt P, Krahmer E (2023) The design and observed effects of robot-performed manual gestures: a systematic review. ACM Trans Hum Robot Interact 12(1):1–62
10. Demir-Lira Ö, Kanero J, Oranç C, Koşkulu S, Franko I, Göksun T, Küntay AC (2020) L2 vocabulary teaching by social robots: the role of gestures and on-screen cues as scaffolds. In: Frontiers in education, vol 5, p 599636. Frontiers Media SA
11. Engwall O, Lopes J, Cumbal R (2022) Is a wizard-of-oz required for robot-led conversation practice in a second language? Int J Soc Robot 14(4):1067–1085
12. Foster P, Tonkyn A, Wigglesworth G (2000) Measuring spoken language: a unit for all reasons. Appl Linguis 21(3):354–375
13. Gordon G, Spaulding S, Westlund JK, Lee JJ, Plummer L, Martinez M, Das M, Breazeal C (2016) Affective personalization of a social robot tutor for children's second language skills. In: Proceedings of the AAAI conference on artificial intelligence, vol 30
14. Han J (2012) Emerging technologies robot assisted language learning. Lang Learn Technol 16:1–9
15. Hockema SA, Smith LB (2009) Learning your language, outside-in and inside-out
16. Hong Z-W, Huang Y-M, Hsu M, Shen W-W (2016) Authoring robot-assisted instructional materials for improving learning performance and motivation in EFL classrooms. J Educ Technol Soc 19(1):337–349
17. Hung I-C, Chao K-J, Lee L, Chen N-S (2013) Designing a robot teaching assistant for enhancing and sustaining learning motivation. Interact Learn Environ 21(2):156–171
18. Iio T, Maeda R, Ogawa K, Yoshikawa Y, Ishiguro H, Suzuki K, Aoki T, Maesaki M, Hama M (2019) Improvement of Japanese adults' English speaking skills via experiences speaking to a robot. J Comput Assist Learn 35(2):228–245
19. Iverson JM (2010) Developing language in a developing body: the relationship between motor development and language development. J Child Lang 37(2):229–261
20. Kanda T, Hirano T, Eaton D, Ishiguro H (2004) Interactive robots as social partners and peer tutors for children: a field trial. Hum Comput Interact 19(1–2):61–84
21. Kanero J, Geçkin V, Oranç C, Mamus E, Küntay AC, Göksun T (2018) Social robots for early language learning: current evidence and future directions. Child Dev Perspect 12(3):146–151
22. Kennedy J, Baxter P, Belpaeme T (2015) Comparing robot embodiments in a guided discovery learning interaction with children. Int J Soc Robot 7(2):293–308
23. Kennedy J, Baxter P, Senft E, Belpaeme T (2016) Social robot tutoring for child second language learning. In: 2016 11th ACM/IEEE international conference on human–robot interaction (HRI), pp 231–238. IEEE
24. Kennedy J, Lemaignan S, Montassier C, Lavalade P, Irfan B, Papadopoulos F, Senft E, Belpaeme T (2017) Child speech recognition in human–robot interaction: evaluations and recommendations. In: Proceedings of the 2017 ACM/IEEE international conference on human–robot interaction, pp 82–90
25. Kersten AW, Smith LB (2002) Attention to novel objects during verb learning. Child Dev 73(1):93–109
26. Kidd CD, Breazeal C (2004) Effect of a robot on user perceptions. In: 2004 IEEE/RSJ international conference on intelligent robots and systems (IROS) (IEEE Cat. No. 04CH37566), vol 4, pp 3559–3564. IEEE
27. Köse H, Uluer P, Akalın N, Yorgancı R, Özkul A, Ince G (2015) The effect of embodiment in sign language tutoring with assistive humanoid robots. Int J Soc Robot 7(4):537–548
28. Kuhl PK, Tsao F-M, Liu H-M (2003) Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. Proc Natl Acad Sci 100(15):9096–9101
29. Lee S, Noh H, Lee J, Lee K, Lee GG, Sagong S, Kim M (2011) On the effectiveness of robot-assisted language learning. ReCALL 23(1):25–58
30. Levelt WJM (1993) Speaking: from intention to articulation. MIT Press
31. Li J (2015) The benefit of being physically present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. Int J Hum Comput Stud 77:23–37
32. Lopes J, Engwall O, Skantze G (2017) A first visit to the robot language café. In: ISCA workshop on speech and language technology in education
33. Lubold N, Walker E, Pon-Barry H (2016) Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion. In: 2016 11th ACM/IEEE international conference on human–robot interaction (HRI), pp 255–262. IEEE
34. Mavilidi M-F, Okely AD, Chandler P, Cliff DP, Paas F (2015) Effects of integrated physical exercises and gestures on preschool children's foreign language vocabulary learning. Educ Psychol Rev 27(3):413–426
35. Mazzoni E, Benvenuti M (2015) A robot-partner for preschool children learning English using socio-cognitive conflict. J Educ Technol Soc 18(4):474–485
36. Powers A, Kiesler S, Fussell S, Torrey C (2007) Comparing a computer agent with a humanoid robot. In: Proceedings of the ACM/IEEE international conference on Human–robot interaction, pp 145–152
37. Randall N (2019) A survey of robot-assisted language learning (RALL). ACM Trans Hum Robot Interact (THRI) 9(1):1–36
38. Rowe ML, Goldin-Meadow S (2009) Differences in early gesture explain SES disparities in child vocabulary size at school entry. Science 323(5916):951–953
39. Schodde T, Bergmann K, Kopp S (2017) Adaptive robot language tutoring based on Bayesian knowledge tracing and predictive decision-making. In: Proceedings of the 2017 ACM/IEEE international conference on human–robot interaction, pp 128–136
40. Tanaka F, Isshiki K, Takahashi F, Uekusa M, Sei R, Hayashi K (2015) Pepper learns together with children: development of an educational application. Humanoids 2015:270–275
41. Tanaka F, Matsuzoe S (2012) Children teach a care-receiving robot to promote their learning: field experiments in a classroom for vocabulary learning. J Hum Robot Interact 1(1):78–95
42. van den Berghe R, van der Ven S, Verhagen J, Oudgenoeg-Paz O, Papadopoulos F, Leseman P (2018) Investigating the effects of a robot peer on l2 word learning. In: Companion of the 2018 ACM/IEEE international conference on human–robot interaction, pp 267–268
43. van den Berghe R, Verhagen J, Oudgenoeg-Paz O, Van der Ven S, Leseman P (2019) Social robots for language learning: a review. Rev Educ Res 89(2):259–295
44. VanLehn K (2011) The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ Psychol 46(4):197–221
45. Wainer J, Feil-Seifer DJ, Shell DA, Mataric MJ (2007) Embodiment and human–robot interaction: a task-based perspective. In:

RO-MAN 2007—the 16th IEEE international symposium on robot and human interactive communication, pp 872–877. IEEE

46. Wellsby M, Pexman PM (2014) Developing embodied cognition: insights from children's concepts and language processing. Front Psychol 5:506
47. Westlund JK, Dickens L, Jeong S, Harris P, DeSteno D, Breazeal C (2015) A comparison of children learning new words from robots, tablets, & people. In: Proceedings of the 1st international conference on social robots in therapy and education
48. Westlund JMK, Dickens L, Jeong S, Harris PL, DeSteno D, Breazeal CL (2017) Children use non-verbal cues to learn new words from robots as well as people. Int J Child Comput Interact 13:1–9
49. Wu W-CV, Wang R-J, Chen N-S (2015) Instructional design using an in-house built teaching assistant robot to enhance elementary school English-as-a-foreign-language learning. Interact Learn Environ 23(6):696–714
50. Yang G-Z, Bellingham J, Dupont PE, Fischer P, Floridi L, Full R, Jacobstein N, Kumar V, McNutt M, Merrifield R et al (2018) The grand challenges of science robotics. Sci Robot 3(14):eaar7650
51. You Z-J, Shen C-Y, Chang C-W, Liu B-J, Chen G-D (2006) A robot as a teaching assistant in an English class. In: Sixth IEEE international conference on advanced learning technologies (ICALT'06), pp 87–91. IEEE
52. Zaga C, Lohse M, Truong KP, Evers V (2015) The effect of a robot's social character on children's task engagement: peer versus tutor. In: International conference on social robotics, pp 704–713. Springer

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Takamasa Iio received a PhD degree from Doshisha University, Kyoto, Japan, in 2012. Then, he has worked at Intelligent Robotics and Communication Laboratories, ATR, Osaka University, and the University of Tsukuba. Currently, he is an associate professor at Doshisha University, Kyoto, Japan. His field of expertise is social robotics. He is interested in how people's cognition and behavior change through interaction with social robots and how human society changes.

Yuichiro Yoshikawa received the PhD degree in engineering from Osaka University, Japan, in 2005. From 2010, he has been an Associate Professor in the Graduate School of Engineering Science, Osaka University. From 2014, he has been a project coordinator of JST ERATO Ishiguro Symbiotic Human–Robot Interaction Project. His research interests include interactive robotics, therapeutic robots for individual with developmental disorders, and cognitive developmental robotics.

Kohei Ogawa received a Ph. D. in system information technology from the Future Universitu-Hakodate, Japan in 2010. He is currently Associate Professor of Department of Engineering at Nagoya University (2019-) and Invited Researcher of Advanced Telecommunications Research Institute (ATR). His research interests are human robot/agent interaction.

Hiroshi Ishiguro received a DEng in systems engineering from the Osaka University, Japan in 1991. He is currently Professor of Department of Systems Innovation in the Graduate School of Engineering Science at Osaka University (2009–) and Distinguished Professor of Osaka University (2017–). He is also visiting Director (2014–) (group leader: 2002–2013) of Hiroshi Ishiguro Laboratories at the Advanced Telecommunications Research Institute and an ATR fellow. His research interests include sensor networks, interactive robotics, and android science. He received the Osaka Cultural Award in 2011. In 2015, he received the Prize for Science and Technology (Research Category) by the Minister of Education, Culture, Sports, Science and Technology (MEXT).