Research on Dialogue-Based CALL

Integrating Tutoring and Implicit Learning:

The Design of an Automatic Joining-in Type Robot Assisted Language Learning

AlBara Jamal Khalifa

September 2019

Abstract

Rapid advances in technology have impacted the field of language learning by providing different ways to support the learning process. The number of research works on computer-assisted language learning systems has increased dramatically over the last 20 years, and there is evidence suggesting that technology-based language learning can be as effective as teacher-delivered instruction. A further step was taken when learners began to engage in dialogues with systems. A more practical environment was created using dialogue-based computer-assisted language learning systems, where a conversation is conducted between the system and a human learner. More technological advances benefited dialogue-based computer-assisted language learning systems, making it possible to offer goal-oriented tasks with virtual embodied agents. Computer-assisted language learning systems provide human learners with a stress-free environment, where they can practice their second language by a convenient and economical self-learning method, with no limitations on the time spent using the system.

The introduction of robots into language learning systems has been highly useful, since it exploits the robot's embodiment to create a lifelike conversational environment. Studies have shown this to be especially helpful in motivating learners to engage in the learning process.

This dissertation presents the design of a novel robot-assisted language learning system that uses two humanoid robots to help learners of English as a second language to improve their practical skills in the language. The joining-in-type humanoid robot-assisted language learning approach uses two robots to conduct a goal-oriented conversation with the human learner in a second language. The system uses implicit learning as the main learning style to teach the usage of a specific expression form. A mix of tutoring and peer learning is implemented in the course of a three-party conversation.

The design of the system helps to regulate the utterances of the human learner through restricting the scope of the utterances to a conversational scenario and through the effect of implicit learning. This effect enables the learner to gain linguistic knowledge, and at the same time it improves the performance of the speech recognition engine.

Keywords

Computer-Assisted Language Learning (CALL)

Dialogue-Based Computer-Assisted Language Learning (DB-CALL)

Robot-Assisted Language Learning (RALL)

Automatic Speech Recognition (ASR)

Automatic Grammar Classification

Implicit Learning

Interactive Alignment

Corrective Feedback

List of Abbreviations

ASR	Automatic Speech Recognition
AGC	Automatic Grammar Classification
AM	Acoustic Model
BLEU	bilingual evaluation understudy
CALL	Computer-Assisted Language Learning
CAPT	Computer-Assisted Pronunciation Training
CMC	Computer-Mediated Communication
CoU	the Correctness of Use measure
DB-CALL	Dialogue-Based Computer-Assisted Language Learning
ICALL	Intelligent Computer-Assisted Language Learning
JIT-RALL	Joining-In-Type Humanoid Robot-Assisted Language Learning
L1	First language
L2	Second language
LM	Language Model
OOV	Out-of-Vocabulary
R1	Teacher robot
R2	Student robot
RALL	Robot-Assisted Language Learning
SDS	Spoken Dialogue Systems
SLA	Second Language Acquisition
WER	Word Error Rate

Contents

1. Introduction **1**

Steps toward robot-assisted language learning systems 1

Computer-assisted language learning 1

Dialogue-based CALL systems 2

Robot-assisted Language learning 3

Language learning issues 4

Tutoring and peer learning 4

Implicit learning 5

Interactive alignment 6

The problem 6

Thesis statement 7

Thesis structure 9

2. Related Works **11**

3. Joining-In-Type Humanoid Robot-Assisted Language Learning System **15**

Introducing the system 15

General experimental setup 17

Why two robots? 20

Wizard of Oz vs automatic system 20

Design of the scenarios 21

Participants 24

Limitations of the system 25

Expected proficiency level of participants 25

Scenarios 26

Automatic robot response and corrective feedback 26

4. Experiments to Evaluate the Implicit Learning Effect in JIT-RALL System **27**

How important is implicit learning? 27

How to calculate the implicit learning effect? 28

Effect of repetitive queries on implicit learning 30

Retention effect 30

Measure used 30
Experiments 31
Effect of implicit learning over two conversations 31
Results 32
Retention effect over four weeks 34
Results 35
Dividing learners by their progress in learning 36
Discussion 37

5. Data Collection in JIT-RALL System 39
Introduction 39
Enhancement of data collection mechanism 40
Experiment 41
Measures used 42
Analysis 43
Discussion 46

6. Introducing ASR into JIT-RALL system 49
Accuracy of ASR with L2 50
Semi-automatic creation of the language model 51
Experiment 52
Automatic grammar classification 53
Results 54
Discussion 56

7. Conclusion and future work 59

Acknowledgments 61

Bibliography 63

Appendix: Automatic Grammar Classification 69

List of Figures

3.1 Front and back view of experimental setup, where the two robots are conducting a conversation with the learner 18

3.2 Conversational structure of JIT-RALL system 22

4.1 Average results using scenario A (past and perfect tenses), showing the effect of implicit learning over two conversations (20 participants) 33

4.2 Average results using scenario B (causative verbs, passive voice, and negative questions), showing the effect of implicit learning over two conversations (37 participants) 34

4.3 Average results in the first week, consecutive training weeks, and retention week, showing the retention effect over four weeks 35

4.4 Average results for learners in the upper half and lower half showing the retention effect over four weeks 37

5.1 Percentage of utterances using CoU in the pre-test and post-test 43

5.2 Percentage of utterances using CoU in all scenarios over the experiment's days. Red line indicates the linear trend of the data, showing improvement over these days 44

5.3 Percentage of utterances of the Difficult Scenarios group and the Easy Scenarios group using the CoU 45

5.4 Percentage of utterances using the CoU in the "repeating" sessions 45

List of Tables

3.1 Example of a conversational scenario. 24

4.1 Order of questions for the control and experimental groups. 29

6.1 Number of utterances using the CoU measure, explained in Chapter 5, relative to the automatic grammar classification (AGC), identified as the first mechanism in the previous section. 55

6.2 Number of utterances using the CoU measure, explained in Chapter 5, relative to the BLEU score classifier. 56

6.3 Number of utterances using the CoU measure, explained in chapter 5, relative to the edit distance classifier. 56

A.1 Examples of expression forms and their tagged forms. 71

Chapter 1

1. Introduction

Rapid advances in technology have affected the field of language learning by providing different ways to support the learning process. Technologies applied to language learning classrooms have included the use of radio, television, audio recordings, and videotapes. The computer was an important technological advance, and not surprisingly its deployment in language learning has been effective and impactful.

Steps toward robot-assisted language learning systems

Computer-assisted language learning

Since the 1980s, computer-assisted language learning (CALL) systems have been adopted to allow learners to practice a second language (L2) through interaction with a computer and thus develop their L2 proficiency (Bibauw et al., 2019). Findings on using CALL obtained in several studies have demonstrated increased achievement in vocabulary acquisition, reading comprehension, and writing skills; however, only a few studies have focused on listening and speaking.

Using computer-mediated communication (CMC) in language learning was considered pedagogically superior to CALL systems, since it includes interaction with an actual human tutor through an e-mail or other type of program (Liu et al., 2002). The performance of the automatic speech recognition (ASR) engine used in CALL systems had about 25% accuracy in recognizing a non-native speaker's speech. This poor performance had an influence on many educators' dismissal of CALL systems at that time.

In the 1990s, more focus was devoted to the design of pedagogically effective CALL activities than to the technological development of these systems, especially after the proposal of Carol Chapelle from Iowa State University that computers should be viewed as a participant that can facilitate communication (Liu et al., 2002). It was later found that students started to prefer learning with a computer program and that their anxiety levels were reported to lessen, which made them more active participants in the learning process. The number of research works on CALL systems has increased dramatically over the past twenty years, and evidence suggests that technology-based language learning can be as effective as teacher-delivered instruction (Bibauw et al., 2019).

Dialogue-based CALL systems

A dialogue-based CALL (DB-CALL) system is a communicative CALL where the "activity consists for the learner to engage in a dialogue with an automated interlocutor in a L2" (Bibauw et al., 2019). It originated with the idea of introducing artificial intelligence techniques to develop a communicative CALL that could implement meaningful conversation with the learner. A group of systems classified as intelligent CALL (ICALL)

followed this trend and focused on correction of written output. Then researchers began, at the end of the 1990s, to analyze and provide feedback on spoken output through computer-assisted pronunciation training (CAPT) systems (Neri et al., 2001; Dalby & Kewley-Port, 1999).

A third category of DB-CALL systems is represented by the spoken dialogue systems (SDS), which focus on dialogue management and emphasize the construction of the conversation itself, such as Galaxy (Seneff et al., 1998). “Let’s Go”, a telephone-based bus schedule information system, is an example of applying the existing SDS framework to language learning (Raux & Eskenazi, 2004). Another example is an interactive computer game that simulates a conversation to help native speakers of English learn Mandarin (Seneff et al., 2007) or native Mandarin speakers to learn English (Seneff et al., 2004). More technological advances benefited DB-CALL systems to offer goal-oriented tasks with virtual embodied agents like SPELL (Anderson et al., 2008), DEAL (Hjalmarsson et al., 2007), and POMY (Lee et al., 2014).

Robot-assisted Language learning

A robot-assisted language learning (RALL) system is a type of DB-CALL system that uses the embodiment of the robot to create a lifelike conversational environment. In what can be considered the first application of a robot to language learning, Kanda et al. developed a robot and introduced it in a classroom. They concluded that “a robot that possesses a humanoid body will be more successful at sustaining interaction because people see it as similar to themselves and that it interacts as they do” (Kanda et al., 2004). In another experiment, Lee et al. (2011) designed a course of English lessons

using robots for Korean elementary school students. They found significant improvement in the students' speaking skills as well as the ability of RALL to improve the students' satisfaction, interest, confidence, and motivation.

The effect of a robot's physical presence on cognitive learning gains was studied by Leyzberg et al. (2012) in comparison with other learning means (i.e., a video of the robot and voice of the robot). They found that participants who were provided advice by the robot directly solved most puzzles faster on average. This effect could be due to the nonverbal modalities that can be offered by robots to an interaction, such as gestures, nodding, and face tracking. These modalities raise the level of experience closer to lifelike communication and provide the learner a more realistic environment. An increasing number of studies have reported that introducing robots enhances learners' interest, motivation, and engagement. This effect was found to be significant in retaining vocabulary in an experiment where a robot simply assisted the human teacher in a second-language classroom (Alemi et al., 2014). They reported that children using the assistive robot retained more vocabulary after two weeks than did the children not using it.

Language learning issues

Tutoring and peer learning

Progress in learning an L2 is believed to be best when using one-on-one tutoring by a skilled instructor, as reported by Bloom (1984) in a review of multiple studies: "...we were astonished at the consistency of the findings as well as the great differences in student cognitive achievement, attitudes and academic self-concept under tutoring as compared

with the group methods of instruction". The one-on-one learning style offers the learner more chances to use the language by communicating more intensively with the teacher. On the other hand, the learner could be exposed to various learning styles in a classroom. Although a classroom often has the problem of having many students who share the opportunity to communicate with the teacher, students could experience different beneficial activities like repeating after the teacher, answering questions, receiving correction sometimes, and learning by listening to interactions between other students and the teacher. They can also be asked to collaborate in more complex tasks and present their thoughts or ideas. Consequently, a combination of tutoring and peer learning is considered effective for learning various aspects of communication in L2.

Implicit learning

Peer learning in a classroom can be considered an implicit learning effect, while tutoring provided by a teacher is mainly an explicit learning effect. Implicit learning is the acquisition of knowledge about the underlying structure of a complex stimulus environment with a spontaneous learning process but without conscious operations (Reber, 1967; Ellis & Bogart, 2007). Explicit learning is a more conscious way of building a structure using the tutoring provided and logic through the exertion of mental effort. "The acquisition of L1 is implicit and is extracted from experience of usage rather than from explicit rules" (Ellis & Bogart, 2007), unlike L2 where implicit learning may not suffice alone, and "[a]dult attainment of L2 accuracy usually requires additional resources of explicit learning" (Ellis & Bogart, 2007). Each kind of learning promotes different aspects of L2 acquisition (SLA).

Interactive alignment

One of the manifestations of the implicit learning effect that can be detected within human conversations is the interactive alignment phenomenon. Within a dialogue, the production and comprehension of the language between the interlocutors become automatically and implicitly aligned on many linguistic levels (Pickering & Garrod, 2004). This interactive alignment phenomenon was detected between participants with the highest proficiency and the lowest proficiency in an experiment on three-party conversations in L2 by Yamamoto et al. (2015). Borrowing each other's linguistic features is a sign of implicit learning, which is "how one develops intuitive knowledge about the underlying structure of a complex stimulus environment" (Reber, 1967), which is the language in this case.

The problem

In order for a CALL system to provide an automatic response as well as automatic corrective feedback to the learner, the ASR engine is a major part that must function properly. Recognizing L2 speech is a challenge even for state-of-the-art ASR engines because it contains various levels of pronunciation quality in addition to lexical, syntactic, and semantic errors. Giving appropriate corrective feedback to each learner is correlated with the performance of the ASR, and it is difficult to consider the wide variety of proficiency of L2 learners and the various reasons that cause the learner to produce erroneous utterances.

Another problem is that few RALL studies have employed adults as learners in the system. Most of the studies were conducted on children, and they offer limited information on the applicability of these systems for adults.

Children acquire L2 in a different manner than adults. Children tend to acquire L2 in a more implicit and unconscious way, which is closer to their acquisition of L1, while adults exert more conscious effort in the process. Another point is that the vocabulary available to children is small, and their understanding is less abstract when compared to adults. These factors result in having higher standards of satisfaction and confidence in mastering the L2 for adults, while for children it is a matter of communicating their needs.

Thesis statement

The main purpose of this work is to develop a joining-in-type humanoid robot-assisted language learning (JIT-RALL) system where two robots conduct a goal-oriented conversation with a human learner in L2. The system uses implicit learning as the main learning style to teach the usage of a specific expression form. Moreover, a mix of tutoring and peer learning is implemented in the three-party conversations.

The three-party conversation comprises a teacher robot (R1) representing a participant with a high level of proficiency, a student robot (R2) representing an intermediate level of proficiency, and the human learner who is expected to have a low level of proficiency. R1 continuously asks questions to both R2 and the human learner. When the question is aimed at the human learner, a tutoring effect is expected, since the interaction is direct between the teacher and the student. When the question is aimed at R2, the human

learner is expected to experience implicit learning through listening to the interaction between R1 and R2. Interactive alignment is expected when R1 asks the human learner to join in the conversation using a similar question that was asked to R2, and thus he/she could have received an implicit learning effect unconsciously from the previous answer of R2 and be able to answer the question.

This work presents a series of experiments using the JIT-RALL system to:

1. Measure the effect of implicit learning in teaching expression forms to human learners
2. Measure the effect of repetitive queries on implicit learning
3. Find the level of retention
4. Propose a prototype of an automatic JIT-RALL system with a corrective feedback mechanism

The main features of JIT-RALL that can be considered additional capabilities over the conventional dialogue-based CALL systems are:

1. The design of the human-robot interaction considers the performance of the ASR engine by restricting the variety of the learner's expression form through showing an appropriate one in the dialogue between two robots. The learner is expected to learn implicitly from the robots' utterances and to become capable of producing grammatically better and expected utterances, which in turn keep the recognition performance high.
2. The design of the conversation's flow promotes an implicit knowledge of using proper expression forms by repeatedly inducing the learner to use the same expression form with different content, which encourages the proceduralization of

the learner's existing knowledge. In other words, the declarative knowledge of the grammar gained using explicit learning evolves into practical implicit knowledge by automating linguistic routines that can be used naturally in the conversation.

3. To handle the difficulties of producing proper expression forms by the learner, the system applies tutoring by showing examples of the dialogue between the two robots when the learner makes an erroneous response.

Thesis structure

This dissertation is organized as follows:

Chapter 2: Related Works

This chapter summarizes some of the previous studies conducted using RALL systems. Related works on the idea of implicit learning and interactive alignment are also presented. Furthermore, the effect of repetition in education with robots is explained.

Chapter 3: Joining-In-Type Humanoid Robot-Assisted Language Learning System

This chapter introduces the JIT-RALL system and discusses the general experimental setup and why introducing robots was helpful to the system. The problem of the poor ASR performance is discussed, as well as how this issue was handled using the Wizard-of-Oz method. Then, the design of the scenario is discussed. Next, information is presented on the participants in the experiments conducted using JIT-RALL. At the end of the chapter, the limitations of the system are mentioned.

Chapter 4: Experiments to Evaluate the Implicit Learning Effect in JIT-RALL System

This chapter describes the effect of implicit learning and how it can be measured. Details on the experiments conducted using JIT-RALL are presented and the results are discussed.

Chapter 5: Data Collection in JIT-RALL System

In this chapter, a discussion is given on the method of data collection using the JIT-RALL system, as well as a way to enhance this method. An experiment using the new approach is described and an analysis of the results is presented.

Chapter 6: Introducing ASR into JIT-RALL system

This chapter presents some ideas on how to introduce the ASR into the JIT-RALL system and how to tackle the poor performance of recognizing L2. Automatic grammar classification mechanisms are also discussed in this chapter.

Chapter 7: Conclusion and future work

This chapter summarizes the main contributions of this dissertation and discusses future work.

Chapter 2

2. Related Works

In what can be considered the first application of a robot to language learning, Kanda et al. developed a robot, named Robovie, and introduced it in a classroom to communicate with Japanese elementary school pupils for three weeks as an English peer tutor. Although the interaction gradually decreased because of the limited patterns of speech synthesis and the limited vocabulary of the ASR, some students continued to interact with Robovie. They concluded that “a robot that possesses a humanoid body will be more successful at sustaining interaction because people see it as similar to themselves and that it interacts as they do” (Kanda et al., 2004).

In another experiment, Lee et al. (2011) designed a course of English lessons using a robot for Korean elementary school students and measured its cognitive effect on oral skills and affective effects. They found a significant improvement in their speaking skills, though not in listening skills. On the affection level, they found that RALL promotes and improves students’ satisfaction, interest, confidence, and motivation.

The effect of pairing a human teacher with an assistive robot was evaluated in an L2 classroom for children (Alemi et al., 2014). Children with both a teacher and a robot learned and retained more vocabulary than did children with only the teacher.

Recently, the use of social robots for teaching a second language to preschool children has been under development in a research project undertaken in Europe (Belpaeme et al., 2015). This project aims to not only teach English to children having other European native languages, as well as Dutch and German to immigrants, but also respond to children's actions and engage with them adaptively while tutoring. One of their papers reported an experiment on how the social behaviors of the tutor robot affected child second-language learning (Kennedy et al., 2016). Although children showed significant improvement between pre- and post- tests under conditions of both high verbal availability of the robot and low verbal availability, the difference between the two conditions reflected no difference in the overall improvement gained.

As for the effect of repetition in education, another paper in the project measured this in collaborative tasks between children and a peer robot (Baxter et al., 2016). The results generally showed positive effects on child performance, and they indicated that this was driven by the individuals because the performance improved even with sparse feedback in the peer-peer interactions.

In terms of implicit learning, borrowing another’s expression while in a dialogue is associated with interactive alignment (Pickering & Garrod, 2004) or entrainment. Interactive alignment is an unconscious process in which interlocutors tend to re-use lexical, syntactic, and other linguistic structures after their introduction. This alignment

was observed in various areas such as lexical choice (Brennan & Clark, 1996), syntax (Reitter & Moore, 2007; Mizukami et al., 2016), and style (Niederhoffer & Pennebaker, 2002). Alignment or entrainment occurs not only in human-human conversation but also in dialogues between a system and a user. Furthermore, Fandrianto and Eskenazi (2012) proposed a dialogue strategy to help regulate users' pronunciation as a way to improve the system's ASR performance.

Chapter 3

3. Joining-In-Type Humanoid Robot-Assisted Language Learning System

Introducing the system

The main purpose of this work is to develop a joining-in-type humanoid robot-assisted language learning (JIT-RALL) system, where two robots conduct a goal-oriented conversation with a human learner in L2. One robot plays the role of a teacher robot (R1) and represents a participant with a high level of proficiency. The second robot plays the role of a student robot (R2) and represents an intermediate level of proficiency. The system assumes the human learner has a low level of proficiency in English.

Tutoring in a one-on-one style is a desirable learning approach, since the expertise of the tutor is dedicated to one student. However, there are also various benefits of the classroom style, where the learner is exposed to peer learning in addition to tutoring, although there is no total focus on a single student in this case. In order to combine the advantages of both styles, three-party conversation can provide a good tradeoff. Tutoring from a teacher robot along with peer learning from a peer robot can be a good

arrangement for integrating the benefits of both styles while maximizing the opportunities to participate.

R1, as the participant with the highest proficiency level, initiates the conversation and continuously asks questions to both R2 and the human learner. When the question is aimed at the human learner, a tutoring effect is expected, since the interaction is direct between the teacher and the student. When the learner has difficulty answering, R1 supports the process using a corrective feedback mechanism. When the question is aimed at R2, the human learner is expected to experience implicit learning through listening to the interaction between R1 and R2. Interactive alignment is expected when R1 asks the human learner to join in the conversation using a similar question that was asked to R2, and thus he/she could receive an implicit learning effect unconsciously from the previous answer of R2 and be able to answer the question. Because this system invites the human learner to join in triad conversations, we called it a joining-in-type humanoid robot-assisted language learning system.

The JIT-RALL system may have an advantage in situations where the human learner finds it stressful to participate in face-to-face interaction. Learners of an L2 are often able to follow a spoken L2 dialogue when listening passively but find it difficult to step in and speak actively. Here, cultural differences need to be taken into account; for example, many Japanese learners of English find it stressful when they are suddenly required to say something in English without having enough time for thinking and preparation (Wilcock & Yamamoto, 2015).

The main features of JIT-RALL that can be considered additional capabilities over the conventional dialogue-based CALL systems are as follows.

1. The design of the human-robot interaction considers the performance of the ASR engine by restricting the variety of the learner's expression form through showing an adequate one in the dialogue between two robots. The learner is expected to learn implicitly from the robots' utterances and become capable of producing grammatically better and expected utterances, which in turn keep the recognition performance high.
2. The design of the flow of the conversation promotes implicit knowledge of using the proper expression forms by repeatedly inducing the learner to use the same expression form with different content, which encourages the proceduralization of the learner's existing knowledge. In other words, the declarative knowledge of the grammar that was gained using explicit learning evolves into practical implicit learning by automating the linguistic routines that can be used naturally in a conversation.
3. To handle the difficulties the learner faces in producing the proper expression form, the system applies tutoring by showing examples of the dialogue between two robots when the learner makes an erroneous response.

General experimental setup

We have developed a prototype joining-in-type RALL system using two NAO robots to explore how to integrate tutoring and implicit learning in SLA. Focusing on a specific type of expression form, the human learner is given an opportunity to construct utterances with appropriate expression forms in lifelike face-to-face conversation. The two NAO

robots are placed standing on a table in front of the human learner as shown in Figure 3.1. The NAO robots do not have facial expressions, so movements of the head and body were the only kinds of body language they exhibited. The robots move their head toward the learner when talking to him/her, or toward each other when the conversation is between them.

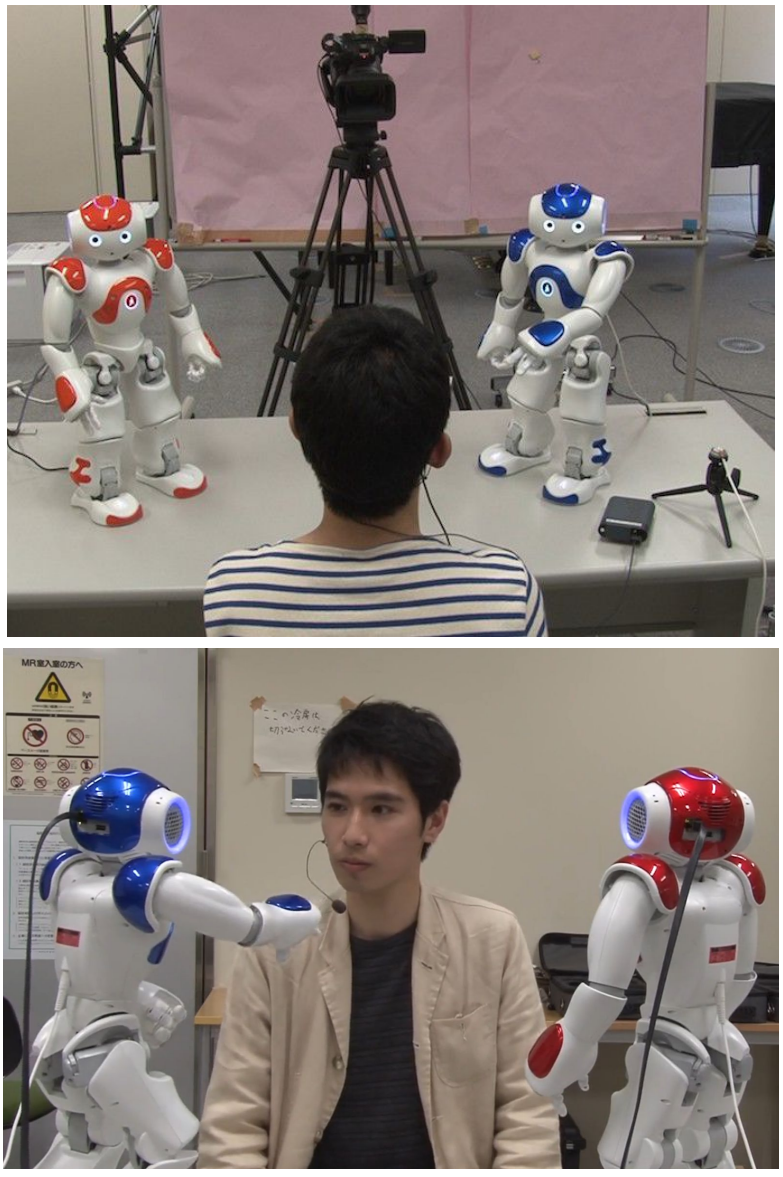

Figure 3.1: Front and back view of experimental setup, where the two robots are conducting a conversation with the learner

In order to build successful interaction between the robot and the human learner, the robot should communicate both verbally and nonverbally. For non-verbal communication, the robot's embodiment feature was exploited by having an expressive set of gestures that were automatically chosen from a library of gestures according to keywords detected in the sentence to be uttered. For this purpose, we used the robot's "AlAnimatedSpeech" module, which is a built-in component of the NAO robot's operating system. For the verbal communication, we used the robot's text-to-speech engine to produce the utterances. The utterances produced by R1 were set to a faster rate than those by R2 to represent the difference in proficiency level between the robots, in addition to the feature of R1 initiating questions throughout the conversation.

The two robots are controlled manually by an experimenter using a Wizard of Oz method. We chose this way to avoid mis-recognition by the ASR system, since L2 is a challenge even for state-of-the-art ASR engines. The experimenter could make the tutor robot repeat the question if the human learner could not answer, say a sample answer and repeat the question, or just pass over the current question and continue the conversation in cases where the learner could not answer at all.

To record the conversation for future analysis, we used two cameras: one in front of the human learner to record his/her activities and facial expressions, and the other behind the human learner to focus on the robots' actions toward the learner. Audio was recorded using a headset microphone worn by the human learner.

The participants were instructed to get involved in the conversation with the two robots by answering their questions, to wait for a while for the robot to repeat it or to give a hint of the answer if the question was not clear, and to speak naturally and clearly.

Finally, after the conversation came to an end, the participants were asked to fill out a questionnaire to measure their attitude toward the experiment. Some questions were about their previous experience using English and dealing with robots. Other questions were about their impressions of the robots. They were also asked to evaluate their interaction in the conversation and to give their overall impression of the experiment. The responses in the questionnaire were given on a seven-level scale.

Why two robots?

The novel approach of the JIT-RALL system in integrating both tutoring and peer learning into a language learning system is based on having a three-party conversation. RALL systems usually use one robot to interact with the learner, sometimes accompanied with use of a tablet computer. This setup offers explicit learning in general, and sometimes implicit learning when a game or a conversation is conducted. JIT-RALL proposes a setup that combines explicit learning through tutoring from a teacher robot and implicit learning through peer observation of the interaction between the two robots.

Wizard of Oz vs automatic system

The recognition of L2 speech is a challenge even for state-of-the-art ASR. A word error rate (WER) of 40% was previously found (Ping, 2008), but this figure dropped to only 39% after a decade (Knill et al., 2019). Due to this formidable difficulty, and to focus on

the pedagogical aspects of the system, we chose to use a Wizard of Oz method to control the experiment's conversation.

The ultimate goal of the system's development is to become a fully automatic language learning system. To reach that goal, we have considered different techniques in the design of the conversation, such as limiting the scope of the conversation. We can also apply the effect of implicit learning throughout the conversation to raise the predictability of the learner's utterances. The interactive alignment that develops from the implicit learning results in expected responses from the learner, which should be considered when designing the language model of the ASR engine.

Design of the scenarios

The conversational scenarios were designed to mimic daily conversations that begin in greetings and move on to chatting on topics like music, movies, sports, travel, new products, and food. These scenarios were designed to draw the human learner into the conversation in a question-and-answer style, and they contained a variety of sentence patterns, such as yes/no questions and 5W1H questions. Figure 3.2 shows a diagram of the conversational structure used in the JIT-RALL system.

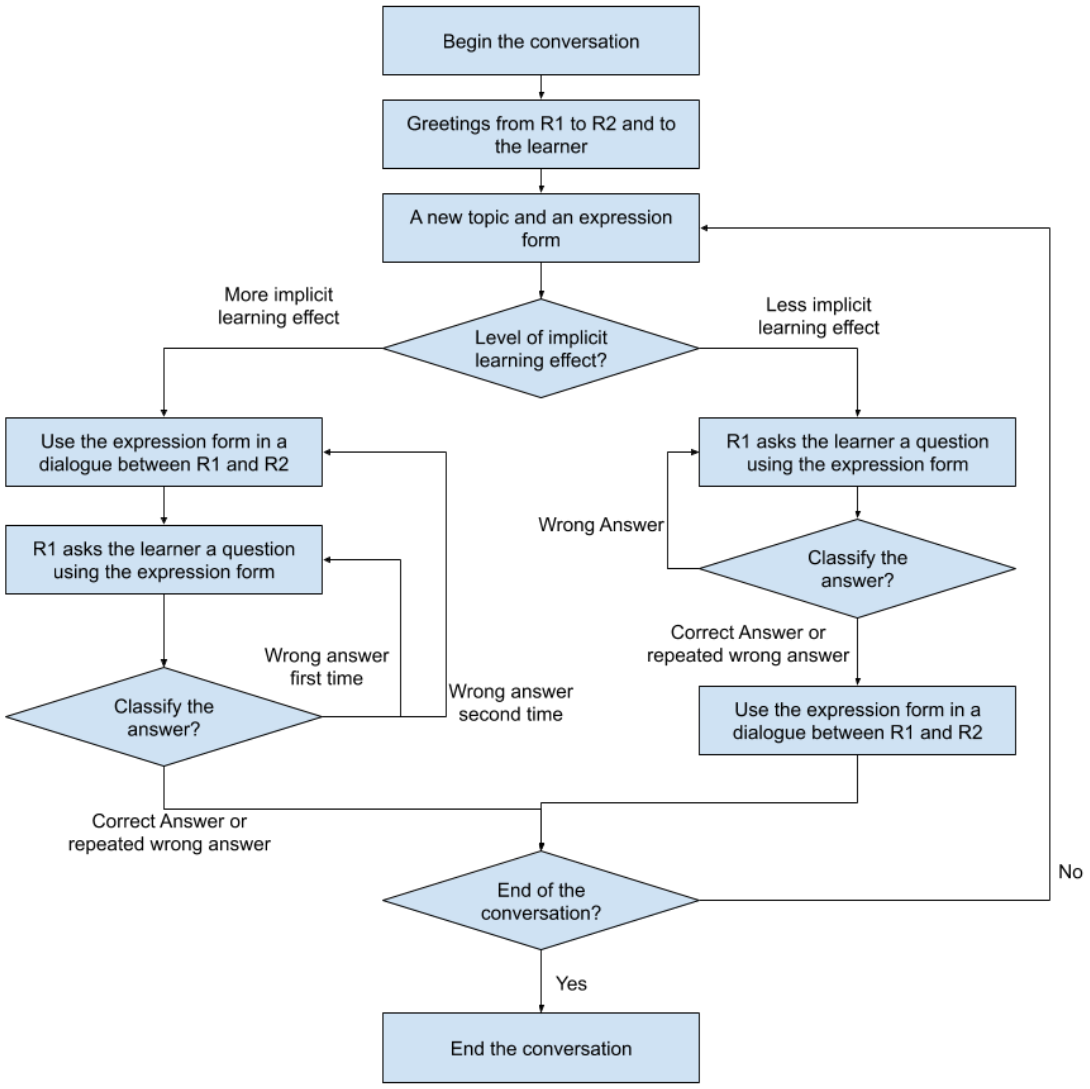

Figure 3.2: Conversational structure of JIT-RALL system

The tutor robot R1 asked the same or similar questions to learners and induced the learner to use the same expression forms as those used by the peer learner robot R2. Some of the questions were expected to be answered in pre-selected expression forms. We used the following set of expression forms in our experiments on the JIT-RALL system:

1. Past, present, and perfect verb tenses

2. Causative verbs
3. Passive voice
4. Answering negatively posed questions
5. Inanimate subjects

Table 3.1 shows a sample of the dialogues extracted from a series of experiments.

The scenarios used in the conversation are predetermined, and the order in which R1 asks the questions influences the extent of the implicit learning effect. For the best expected level of implicit learning effect, R1 asks a question to R2; after R2's answer to the question is produced, R1 asks the same question to the human learner. The answer expected from the learner would be affected by and aligned to the answer of R2, and this could be considered an implicit learning effect.

Besides having an experimental group, who are being exposed to the implicit learning effect, having a control group is important for establishing a baseline of the effect. To set the lowest possible implicit learning effect while also keeping the same amount of exposure to the utterances from both robots, we suggested changing the order of the questions asked by R1. In the case of the control group, R1 asks the human learner the question first, then asks R2. Over multiple sessions, the amount of knowledge presented to both groups is the same, but a larger effect of implicit learning is expected for the experimental group.

Table 3.1: Example of a conversational scenario.

No.	Speaker	Listener	Utterance
1	R1	R2	What do you think your mother will be given by your father for her birthday present?
2	R2	R1	I think my mother will be given a necklace by my father.
3	R1	R2	That's great.
4	R1	Learner	What were you given for a birthday present last year?
5	Learner	R1	I was given a coffee maker.
6	R1	Learner	Good.
7	R1	Learner	What were you taught by your mother?
8	Learner	R1	I was taught how to cook by my mother.
9	R1	Learner	I see.
10	R1	R2	What were you taught by your father?
11	R2	R1	I was taught how to drive by my father.
12	R1	R2	Good.
13	R1	R2	While driving, I need a car navigation system.
14	R1	R2	Because my car doesn't have a car navigation system.
15	R1	Learner	Don't you have a car?
16	Learner	R1	Yes, I have a car
17	R1	Learner	I see.
18	R1	R2	Don't you have a car?
19	R2	R1	No, I don't have a car.
20	R1	R2	OK.
21	R1	Learner	If you had a car and your car were broken, what would you do?
22	Learner	R1	I would call my father.
23	R1	Learner	Good.

Participants

We recruited 80 Japanese university students between the ages of 18 and 24 as participants in a series of experiments. The participants had acquired Japanese as their L1 and had learned English as their L2 at various proficiency levels; this variety of participants' linguistic profiles was intended to evaluate the effectiveness of the JIT-RALL system for a realistic range of users. More than half of the participants took the Test of English for International Communication (TOEIC), and their scores ranged from

320 to 980, with an overall average of about 564 (990 being the highest attainable score).

Limitations of the system

The JIT-RALL system is limited by the type of learners expected, the variety of topics to be discussed in conversations, the level of recognizing learners' utterances, and the quality of the corrective feedback given to the learner.

Expected proficiency level of participants

The system is designed for Japanese who are learning English as an L2. The focus of the system is to improve the grammatical knowledge of the learner by speaking, rather than through reading or writing. This should affect both their listening and speaking skills, since these are the means to achieve the main goal. However, the English used and expected to be uttered by the learner is simple, which should help them to practice the intended expression form. Consequently, this system may not be appropriate for those who lack basics in English conversation or, at the same time, those who have an intermediate or higher level of English proficiency.

Japanese learners of English are expected to use the system, since the design considers their L1 background. The expression forms used in the scenarios are chosen because Japanese learners often find them difficult. Although the system does not yet use ASR, the acoustic model (AM) and the language model (LM) for the ASR engine that is now being built are trained on Japanese utterances. Therefore, L1s other than Japanese may not work well in this setting.

Scenarios

The JIT-RALL system in its current state is not an automatic system due to the difficulty of applying an ASR engine. The recognition of L2 speech is a challenge even for state-of-the-art ASR. Because of this difficulty, and to focus on the pedagogical aspects of the system, we chose the Wizard of Oz approach to control the conversation. The scenarios in this work had to be predetermined. The conversation is limited by what is in the scenarios, and so the learner's answers would not be handled properly if the conversation were outside of the scenario's structure.

Automatic robot response and corrective feedback

The responses of the system through the robots depend on the Wizard of Oz control by the experimenter. Moreover, the decisions made by the experimenter are limited by the design of the scenario. A response is limited to either repeating the question, if the human learner could not answer it, using a group of corrective feedback techniques like saying a sample answer and repeating the question, or repeating the previous conversation of the two robots and then asking the learner the question again. In the worst case, the experimenter can just pass over the current question and continue the conversation when the learner cannot answer at all.

Chapter 4

4. Experiments to Evaluate the Implicit Learning Effect in JIT-RALL System

How important is implicit learning?

In language learning, implicit learning is the acquisition of knowledge about the underlying structure of the language. The ease we enjoy in producing our L1 comes from the fact that it was built unconsciously, without making a conscious effort to understand and apply the structure of the language. Children in their early years can use language, but they cannot explicitly recognize grammar, syntax, or other linguistic properties. In the case of L2, adults usually require additional resources using explicit learning. This explicit instruction is supported through different means such as books, courses, or several kinds of CALL systems. Practicing a language for use in real life requires one to apply this explicit knowledge to natural and unconscious use. That is why using the language in daily life (i.e., in a different country) can be the best way to practice it, which is mainly an implicit way of learning. The JIT-RALL system offers an opportunity to practice a language in a lifelike, conversational style. It promotes learning of the language in an implicit manner.

Defining the term “implicit learning” and drawing a clear boundary between it and “implicit knowledge” or “implicit memory” is still a subject of some controversy. Explaining such distinctions in detail is out of the scope of this work. To simplify the terms in this paper, we use the factor of attention or consciousness exerted in using the language as the main difference between implicit and explicit learning. Explicit learning is considered learning a grammatical rule explicitly and then using that knowledge to create sentences that implement this learned rule. On the other hand, implicit learning is considered learning a grammatical rule through listening repeatedly to sentences created using that rule, without conscious and explicit awareness of that rule’s structural properties, and then creating sentences while giving attention to the situation rather than to the correct application of the grammatical rule.

How to calculate the implicit learning effect?

The way the scenario is designed in the JIT-RALL system provides the learner with a combination of explicit and implicit learning through tutoring and peer learning, respectively. The order in which R1 asks the questions, to R2 first and then to the learner, or vice versa, changes the amount of implicit learning effect that the learner can obtain. For the best expected level of implicit learning effect, R1 asks a question to R2; after the answer of R2 to the question is produced, R1 asks the same question to the human learner. The answer expected from the learner would be affected by and aligned to the answer of R2, so this can be considered an implicit learning effect.

Besides having an experimental group, who are exposed to the implicit learning effect, having a control group is important to establish a baseline of the effect. To set the lowest

possible implicit learning effect while keeping the same amount of exposure to both robots' utterances, we decided to change the order of the questions asked by R1. In the case of the control group, R1 asks the human learner the question first and then asks R2. Over multiple sessions of the conversation, the amount of knowledge presented to both groups is the same, but a greater effect of implicit learning is expected for the experimental group.

To calculate the effect of implicit learning on the experimental group under equal conditions with the control group, the following question order was devised. During the conversation, the tutor robot R1 asks each learner a group of questions, among them five questions that should be answered in a certain expression form. In the case of the experimental group, three of the questions are asked to R2 first and then to the learner as a training set. For the control group, these same three training questions are only asked to the learner. This means more training will be offered to the experimental group. The two remaining questions are asked to the learner first and then to R2 in both groups as an evaluation set of questions. This procedure is summarized in Table 4.1.

Table 4.1: Order of questions for the control and experimental groups.
(Qs: questions)

Kind of question	**Order of questions from R1 to R2 and to human learner**	
	Control group	**Experimental group**
Training	3 Qs to learner only	3 Qs to R2 first and then to learner
Evaluation	2 Qs to learner first and then to R2	2 Qs to learner first and then to R2

Effect of repetitive queries on implicit learning

The effect of learning is enhanced by repetition (Baxter et al., 2016). The design of the scenarios considers this finding and applies it throughout the experiments. Each expression form is used in every conversation in more than one question. In every conversation, the tutor robot R1 asks each learner a group of questions, among them five questions that should be answered in a certain expression form. Each conversation is repeated once more with the learner. In the repeated conversation, the questions used are almost the same but some words are changed. The goals are to enhance the vocabulary of the conversation and to reduce the effect of short-term memory, which could produce answers from memory instead of by applying the expression form.

Retention effect

To confirm that the implicit learning gained by the learner can be retained over some time, the training session of the two conversations is repeated several times with a period of one week between the sessions. A final repetition is conducted after a break of a week or more from the training sessions. The level of learning effect could thus be measured on different weeks to determine whether the expression forms learned could be retained after a significant amount of time.

Measure used

Measuring the correctness of the learner's answer was done from a grammatical perspective while focusing on the expression form used in the scenario. In this process, we manually evaluated every answer using the following five-level grading:

1. This is the worst grade, and it is assigned to an answer if the learner did not respond or said something like “I don’t know”
2. An answer with wrong expression form
3. An answer using the correct expression form but with wrong semantic content
4. An answer using the correct expression form with correct semantic content but with a grammatical mistake
5. An answer that is grammatically and semantically correct

Experiments

Effect of implicit learning over two conversations

The purpose of this experiment is to evaluate to what extent learners refer to or mimic utterances by the peer learner R2 at first glance (Khalifa et al., 2017; Khalifa et al., 2018; Khalifa et al., 2019). Learners are divided into two control group and an experimental group, and each learner is asked to join a triad conversation along with the two robots. During the conversation, the tutor robot R1 asks the participant five questions that should be answered in a certain expression form. Three of the questions are spoken to R2 and then to the learner in the experimental group, and the same three questions are only spoken to the learner in the control group. The two remaining questions are spoken to the learner first and then to the peer learner robot R2 second in both groups to evaluate the effect of implicit learning under equivalent conditions. This procedure is summarized in Table 4.1.

Every learner performed two consecutive 10-minute conversations, taking a break of 5 minutes between them. The first and second conversation in the session are conducted

based on two kinds of similar scenarios in which the tutor robot R1 asks questions that should be answered in the same expression form but with different vocabulary according to each question.

We designed two kinds of scenarios for the two experiments. For the first experiment, we selected a scenario (scenario A) in which the tutor robot R1 asks questions that should be answered in two different expression forms: the past tense and the present perfect tense, which Japanese learners often mistake in speaking English although they study these forms at a relatively early stage of grammar learning. As the scenario of the second experiment (scenario B), we selected causative verbs, passive voice, and negative questions, which are expected to be difficult for Japanese learners of English. The test procedure is the same as that for the experiment using scenario A. The numbers of participants are 20 for scenario A and 37 for scenario B.

Results

Figures 4.1 and 4.2 show the experimental results as the average rate of using the same expression forms as those used by the peer learner robot R2 for all answers. This ratio shows to what extent the participants refer to or mimic the utterances by the peer learner R2. The results increased by about two times from the first conversation to the second in both experiments using scenarios A and B, which shows that implicit learning generally helps human learners to construct utterances with appropriate expression forms. Results were larger for the learners in the experimental group than for those in the control group, which shows that the more training a question is given, the more effective human learners can construct utterances with adequate expression forms, and the more

repetitive learning is indispensable to making implicit learning effective. The result in the second conversation of the experiment of scenario A (use of past and perfect tenses) is higher by about two times than that of scenario B (use of causative verbs, passive voice, and negative questions), which suggests that the more difficult the expression form used, the more insufficient the training provided is for human learners to learn how to construct the utterance appropriately.

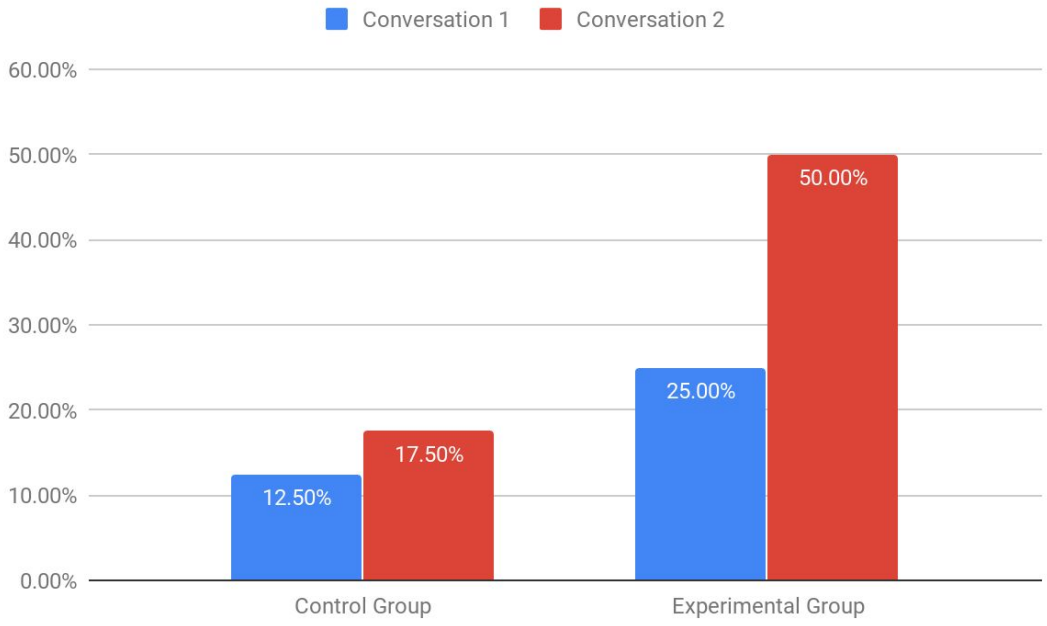

Figure 4.1: Average results using scenario A (past and perfect tenses), showing the effect of implicit learning over two conversations (20 participants)

Analyses of collected utterances of the participants show that many utterances are answers of a single word or a single phrase, or irrelevant answers. Single-word answers suggest that the participant understands the conversational flow but could not express his or her answer in the appropriate expression forms. The irrelevant answers may have been caused by the participants of low proficiency not understanding the conversational flow. The number of irrelevant answers is larger in the experiment using scenario B than in that using scenario A.

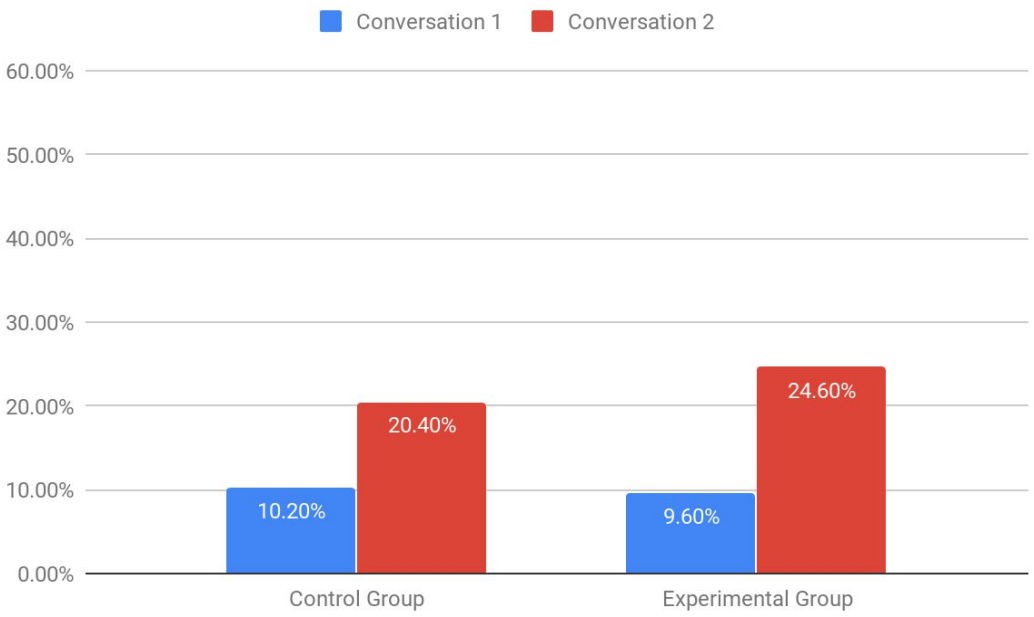

Figure 4.2: Average results using scenario B (causative verbs, passive voice, and negative questions), showing the effect of implicit learning over two conversations (37 participants)

Retention effect over four weeks

In another experiment (Khalifa et al., 2018; Khalifa et al., 2019), we created four different scenarios that were used over four consecutive weeks and in a retention test to evaluate the repetitive effect and the retention of implicit learning. We selected causative verbs and inanimate subjects as the expression forms to be used in the questions, which are expected to be difficult for Japanese learners of English because such an expressional style is far different from that of their native tongue. Weekly sessions of two conversations were held, and each learner was asked 10 questions in every conversation. Learners were divided into a control group and an experimental group, and each participant of both groups had the same number of opportunities to hear the answers from R2. For all 10 questions that were asked to the learners, the tutor robot R1 asked a question to the learner first and then asked the same or a similar question to the

peer learner robot R2 for the control group. In the case of the experimental group, six questions out of ten were asked first to the peer learner robot R2 and then to the learner as shown in Table 4.1. The scenarios of the first conversation of week 1 and week 3 were used in the retention test. We recruited 23 learners from the same population of Japanese undergraduate or graduate students. Among them, 16 could participate in all sessions.

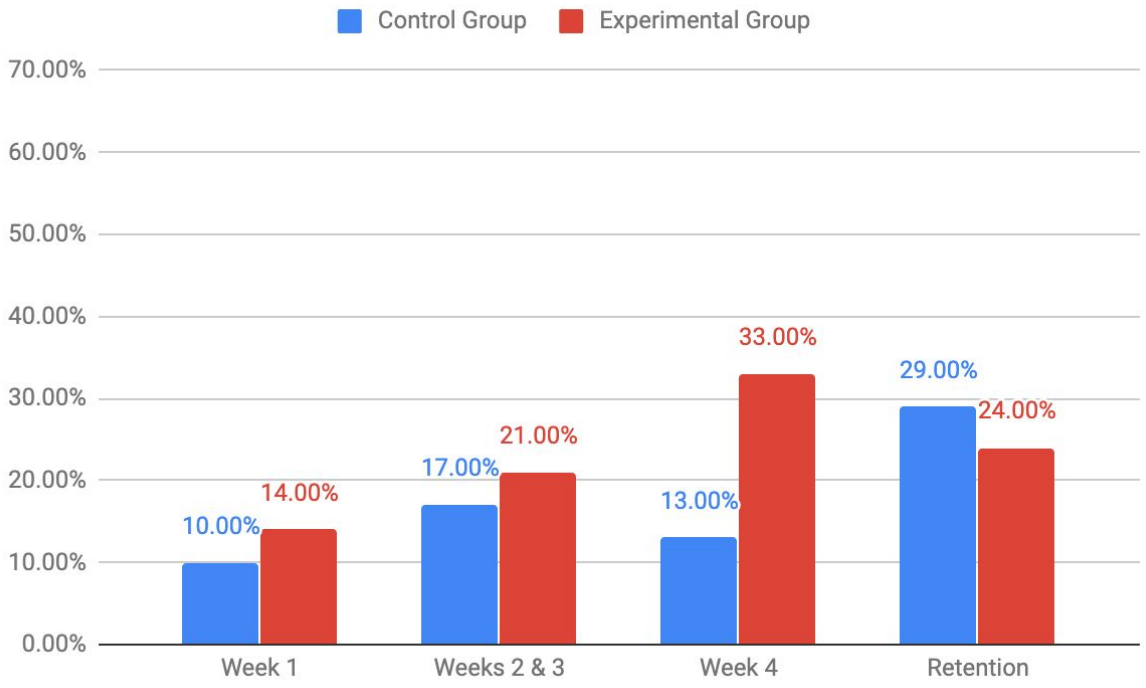

Figure 4.3: Average results in the first week, consecutive training weeks, and retention week, showing the retention effect over four weeks

Results

Figure 4.3 shows the average results in week 1, in the training sessions of three consecutive weeks (intermediate two weeks and final week), and in the retention test. The experimental results increased from the first week to those on the final week. They were larger for learners in the experimental group than for those in the control group.

The difference in the results between the groups of learners is lower for the retention test than for the training sessions. The results in all cases were less than 35%, even for the participants in the experimental group, which indicates a low level of improvement. The next section explores this issue in detail.

Dividing learners by their progress in learning

Implicit learning generally helps human learners to construct utterances with appropriate expression forms; however, the improvement obtained through training over four consecutive weeks does not seem so high. To explore the reason for this, we investigated the improvement of each learner. Figure 4.4 compares improvement in the results for the learners by dividing them according to their results into the upper half and lower half. The achieved result depends on each learner, and its variation is very large as shown in the figure. The achieved result is largely divided into two classes, and the final achievement is 47% for one class and only 3% for the other class. These results suggest that the scenario may be too difficult for learners of the lower-achieving class. Implicit learning is effective for human learners, but mainly for learners of relatively higher proficiency. The teaching material and the conversational scenarios in JIT-RALL systems should be designed to fit the proficiency of the human learner, and a method to measure the proficiency of each learner should be explored.

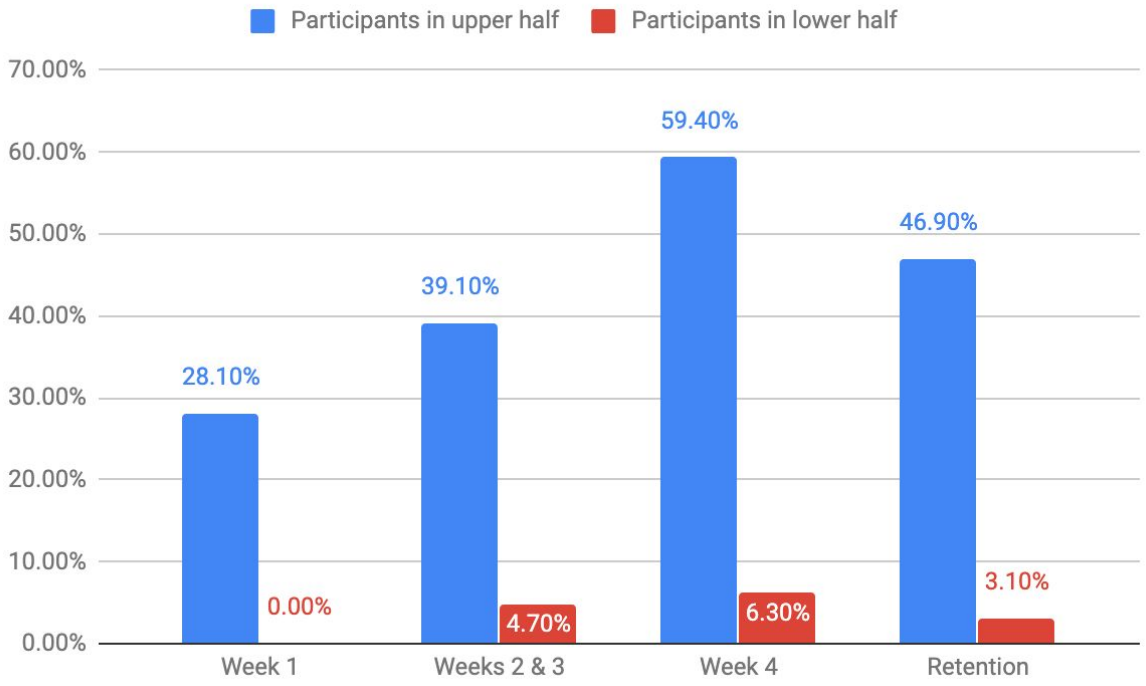

Figure 4.4: Average results for learners in the upper half and lower half showing the retention effect over four weeks

Discussion

The experimental results of the effect of implicit learning over two conversations show that the implicit learning incorporated in the JIT-RALL system is promising for giving learners the chance to obtain the ability to use appropriate expression forms in real conversational situations. The analysis results of the experiment under scenarios A and B show that the human learner used appropriate expression forms at a high ratio when he/she responded just after hearing the answers of the peer learner robot R2, compared with the responses without such training using the answers by R2. These results suggest an effect of implicit learning in the JIT-RALL system.

The experimental results on the retention effect over four weeks show that the repetitive training of implicit learning increases the ratio of using appropriate expression forms.

The ratio is higher for learners in the experimental group than in the control group during the training session; on the other hand, the difference between the ratios for learners of the two groups decreased in the retention test. These results suggest that post-presentation of the answers by R2 for the evaluation questions operates as implicit learning as well as the training questions, especially for retention.

The retention test results of the learners in the experimental group were lower than those in the control group (Figure 4.3); this was unexpected, especially after a significant increase was achieved during the training sessions. The reason could be due to the effect of short-term memory associated with the answers presented by R2 right before the learner's answer. This means that the steep increase in their results in the training session could be a combined effect of short-term memory and implicit learning.

Chapter 5

5. Data Collection in JIT-RALL System

Introduction

One of the main goals of collecting the conversational data in the experiments was to detect the effect of implicit learning using the JIT-RALL system. Different mechanisms were applied to magnify the effect so that it could clearly appear in the utterances of the learners when compared to the control group. The question-ordering mechanism, explained above, played a major role in creating different situations for the two groups. It helped to promote more implicit learning in the experimental group while exposing the two groups to the same amount of knowledge in order to allow a fair comparison between them. Repeating the expression multiple times in the conversation using different content was another mechanism applied to promote implicit learning in the learner. This effect of repetition was even stronger when used over multiple sessions over weeks of the experiment.

The results of analyzing the collected data indicate operation of an implicit learning effect. Consequently, learners of an L2 can use the JIT-RALL system to improve their practical knowledge of using the expression form they have trained on.

Despite the promising results found, the data collection using the JIT-RALL system showed some problems. Assigning learners to the experimental group and to the control group was a random process that did not consider the variety of participants' proficiency levels. Some of the learners faced difficulty in holding a conversation with the robots, which caused their utterances to affect the overall results. They did not show improvements in uttering proper expression forms over the multiple sessions conducted and did not seem capable of understanding the conversation between R1 and R2. With the ultimate goal of creating an automatic JIT-RALL system, the ASR will have difficulty in recognizing such utterances, which will degrade the performance of the system. The corrective feedback provided by the system depends mainly on the performance of the ASR. Poor performance of the ASR will result in unexpected corrective feedback, which will cause the learner confusion.

Another problem is the time needed to conduct the experiments, which made it difficult to recruit more learners. The training of the learners over multiple weeks for the sake of repeatedly using the expression forms with different content was helpful, but it caused the experiment to spread over several weeks. Finding learners who could participate every week became much more difficult than recruiting for a shorter experimental period.

Enhancement of data collection mechanism

We designed an experiment to tackle the above issues. The problem of having a diversity of learners' proficiency levels can be approached by providing more training to those with a low level of proficiency and by offering training material that is appropriate

to the learner's level. The training included "repeating" sessions, where the learners have the opportunity to repeat the robots' utterances instead of participating in conversations. The idea is to raise their familiarity with the robots' way of producing speech and enhance their production of utterances in the conversation. Furthermore, the problem of needing a long time to conduct the experiment was solved by setting up a one-week experiment. In this experimental design, the learner needs to participate in a session every day (for six days) throughout the week as detailed in the following section.

Experiment

Each learner is asked to participate in a three-party conversation with the two robots on a daily basis for six days. A set of scenarios was prepared for use in the conversations over these days. The conversations in the first and sixth days were conducted using the same scenarios in order to use the results of these days as a pre-test and a post-test. The conversations on the other days were considered training for the learner.

Two groups of scenarios were prepared. A "difficult" group contained scenarios that use causative verbs and inanimate subjects as the expression form repeated over the entire conversation. An "easy" group contained scenarios that use past and present tenses for the verbs. The level of difficulty of the scenario is measured according to which expression form is considered difficult or easy for the Japanese learners of English as L2.

During the conversation, the questions were asked by R1 in all cases. R1 asks R2 a question, then asks the question again to the learner. On the first day, all learners were

exposed to one scenario from each group as a pre-test. On the second and third days, a "repeating" session is conducted. In these sessions, a conversation is conducted between R1 and R2. Whenever R2 makes an utterance, the learner is asked to repeat the answer of R2. According to the capability of the learners to perform well in the "repeating" sessions, they are assigned to one of two different groups. The first group, those who did well in the "repeating" sessions, use the difficult group of scenarios on the fourth and fifth days of the experiment, while the others use the easy group of scenarios. The idea is to keep the difficulty of the scenario consistent with the learner's level of proficiency. On the last day, the same scenarios used on the first day were used again as a post-test.

Measures used

Measuring the correctness of the learner's answer was done from a grammatical perspective. We manually evaluated every answer using the following five-level grading:

1. For no answer or an "I don't know" kind of answer
2. For wrong answers (maybe due to misunderstanding of the dialogue)
3. For correct grammar but not the expected expression form
4. For the expected expression form but with some mistakes
5. For the correct and expected expression form

Only the utterances graded with a score of four or five were considered correct answers in the results. Here, this measure is referred to as the correctness of use (CoU) measure.

The word error rate (WER) was used to measure the performance of the ASR in recognizing utterances. The more expected the answer is when compared to the expression forms used in the conversation, the more accurate the recognition of the utterance will be. We used Kaldi as the ASR engine. The acoustic model was trained using the ERJ corpus, which is the English Speech Database Read by Japanese Students.

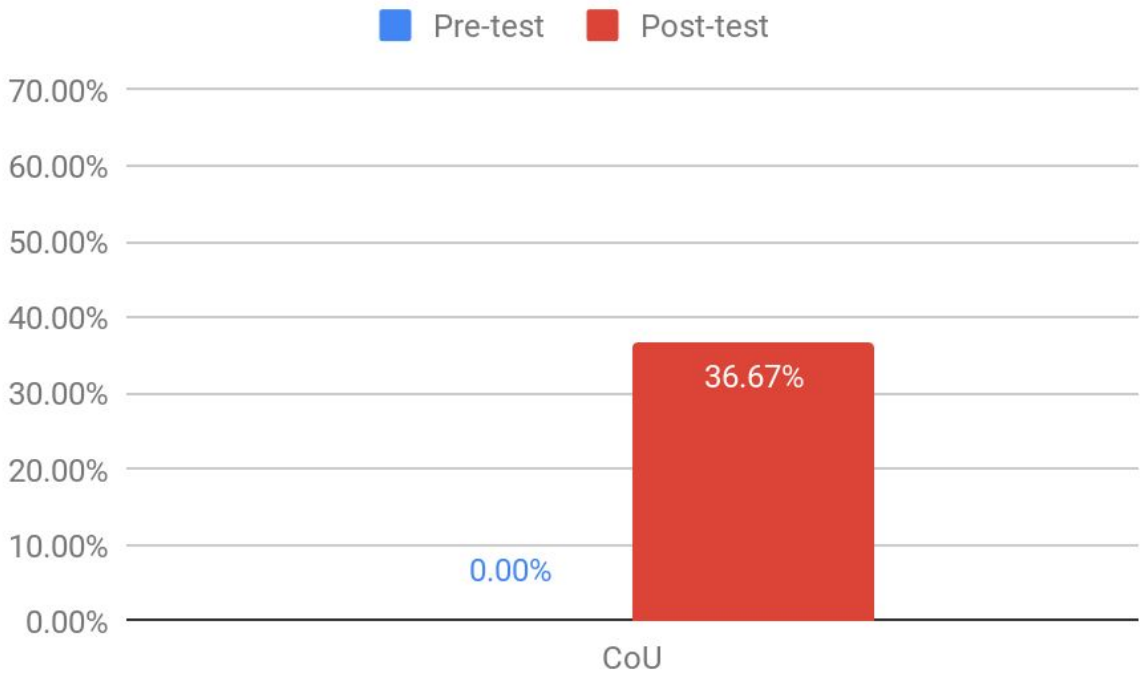

Figure 5.1: Percentage of utterances using CoU in the pre-test and post-test

Analysis

In this analysis, I discuss the implicit learning effect on the learner to utter a grammatically correct utterance. I also discuss the ASR performance in the offline recognition of the recorded utterances.

The data collected for six learners over a period of six days were 358 utterances. Figure 5.1 shows the percentage of the utterances using CoU in the pre-test and the post-test. The learners in this case show a clear improvement in learning the expression form. Figure 5.2 shows the percentage of utterances using CoU in all scenarios over the experiment's days, indicating an overall improvement. There is a change in the rate of improvement after the third day, which could be due to each learner being assigned to a different group of scenarios according to their performance in the "repeating" sessions (Figure 5.4), which caused some difficulty for those with the difficult scenarios (Figure 5.3). Assigning learners to different groups was important for maintaining appropriate levels of difficulty between the "too difficult" and the "too easy" groups.

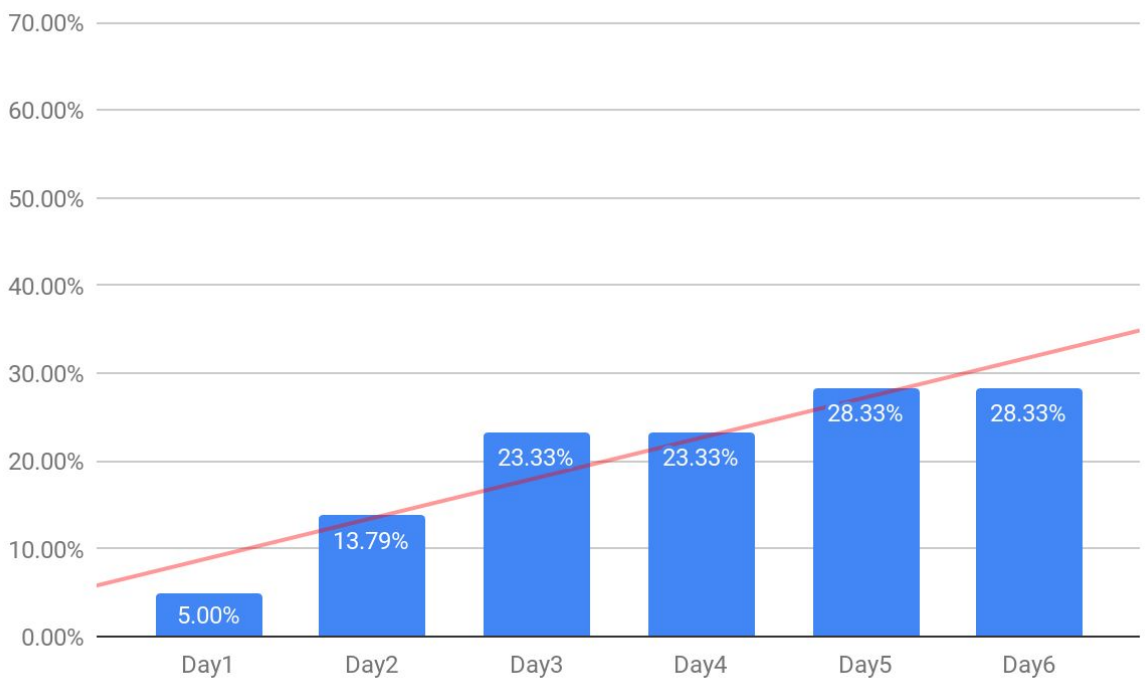

Figure 5.2: Percentage of utterances using CoU in all scenarios over the experiment's days. Red line indicates the linear trend of the data, showing improvement over these days

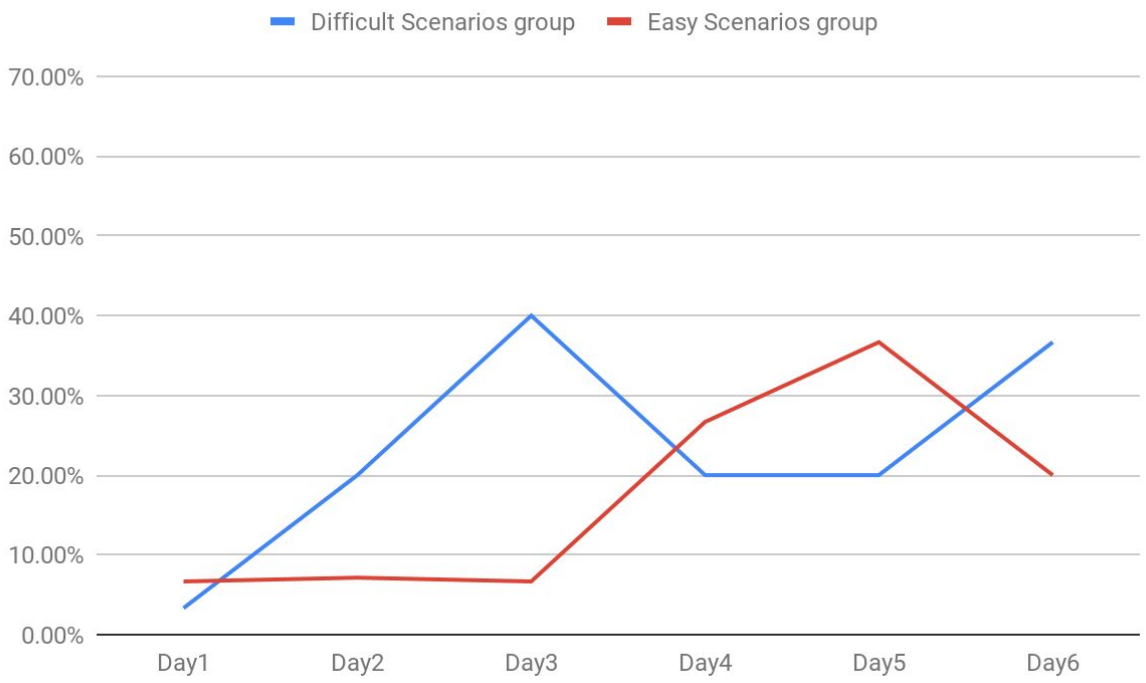

Figure 5.3: Percentage of utterances of the Difficult Scenarios group and the Easy Scenarios group using the CoU

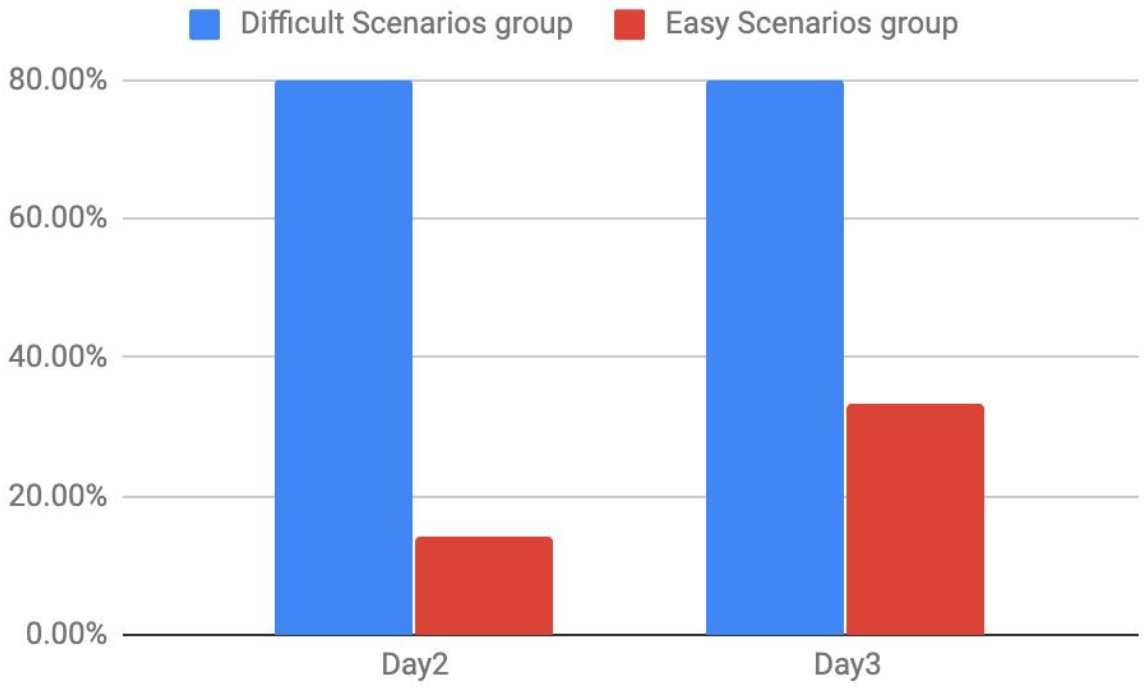

Figure 5.4: Percentage of utterances using the CoU in the “repeating” sessions

For the ASR performance, we used the transcription of the learners' utterances as a corpus to train the LM, in order to find the best performance that the ASR could achieve. The WER was 53.23% (performance of 46.77%). This result is close to that of another experiment we conducted previously with four learners (217 utterances) in a very similar environment and setup, where we had WER = 47.81% (performance of 52.19%). When we analyzed the produced speech we found that the fillers may have been a major cause of the poor ASR performance. Fillers are the sound produced by the speaker in times of hesitation used to signal to others the desire to pause for thinking and to delay yielding his/her turn in the conversation. To confirm this, we modified the speech data by removing all of the fillers from the audio files. The WER was reduced to 27.21% (performance of 72.79%). This means that the fillers degrade performance by about 20%.

Discussion

The results show an improvement in gaining grammatical knowledge through implicit learning in a relatively short time and with relatively better results in comparison to the experiments described in Chapter 4. This indicates that using this experimental setup with more learners could result in a better implicit learning effect. In addition, more learners could be recruited since the time of the experiment was shortened, without losing the benefit of the repetition factor in the conversation.

Another point to note is that the design of this experiment considered the different levels of proficiency of the learners Those with a low level of proficiency, according to the "repeating" sessions, were assigned to a more suitable scenario. If we consider the

group with the “difficult” scenarios and the group with the “easy” scenarios in this experiment comparable to the upper half and the lower half of the experiment described in Chapter 4, then Figure 5.3 demonstrates an improvement in the experimental design over that used for Figure 4.4. The two groups had opportunities for improvement over the days depicted in Figure 5.3, and they seemed to be encouraged in a balanced manner. On the other hand, the results in Figure 4.4 clearly shows a degree of discouragement for one group as well as unbalanced opportunities for learning, since one group (the lower half) apparently had a very difficult scenario for their proficiency level.

Regarding ASR performance, one previous work (Ping, 2008) found the WER for non-native speakers to be about 40%. That was in 2008. A 2019 publication (Knill et al., 2019) showed a figure of about 39% for low-proficiency speakers. Consequently, it is still a challenge for ASR to recognize L2 speech.

However, an appropriate experimental design could help to improve ASR performance. Restricting the scope of utterances of the learner is a major feature of the JIT-RALL system. As mentioned in the analysis section, we have achieved a WER result of 27.21%. Although the utterances could be successfully restricted by the system, achieving that result was possible when the fillers were removed from the speech. Using ASR online during the conversation is the ultimate goal of the JIT-RALL system. In order to reduce the number of fillers produced by the learner, one solution might be to ask him/her to repeat their answer a couple of times.

The ASR results presented in this chapter are based on an LM that was trained using the transcription of the conversation. Designing an LM that could be used in an online manner during a conversation would need to include the possibility of handling the learner's answers according to the scenario used. In the next chapter, we introduce an idea for how to achieve this goal.

Chapter 6

6. Introducing ASR into JIT-RALL system

The ultimate goal in developing the JIT-RALL system is to make it an automatic system that can provide a language learning environment with a corrective feedback mechanism suitable for the implicit learning style. However, the performance of ASR in recognizing an L2 is still a challenge, which affects the corrective feedback accuracy. To compensate for this weakness, the system in its current form uses a Wizard of Oz method to direct language learning without depending on the ASR engine, since a human experimenter controls certain actions of the robots. Keeping in mind the goal of incorporating an ASR engine in the system, we implemented techniques that were helpful both from a pedagogical perspective and for building the capability to regulate the learners' utterances in a way that improves ASR accuracy. We could do this by restricting the scope of the utterances to the conversation's scenario and implementing a system-initiated conversation where R1 asks all the questions. Another important factor in regulating the learner's utterances was the implicit learning effect implemented in the conversation, which is intended to induce the learners to align their utterances to those of R2. This in turn can help to raise the predictability of the utterances and improve the performance of the ASR.

In this chapter, I propose a way that takes a step toward implementing an ASR in the JIT-RALL system through creating a semi-automatically generated corpus that can be used to train the LM. I will also discuss preliminary results of applying different techniques for automatic grammar classification of utterances.

Accuracy of ASR with L2

Using ASR to recognize an L2 is a difficult task, even for state-of-the-art ASR engines, because the utterances contain various levels of pronunciation quality in addition to lexical, syntactic, and semantic errors. Giving appropriate corrective feedback to each learner is correlated with ASR performance, and this is difficult considering the wide variety of proficiency of L2 learners and the various reasons that cause the learner to produce erroneous utterances. The mechanisms applied in the JIT-RALL system could help to raise the level of correctness of the utterances produced by the learner, which in turn should be a helpful step in developing the ASR engine.

The JIT-RALL system could successfully restrict the scope of the learners' utterances. However, the ASR engine should have an LM that considers more possibilities of the vocabulary that could be used in the conversation's expression forms. One of the factors that affect the ASR performance is the LM's out-of-vocabulary (OOV) words. Most ASR systems are closed-vocabulary recognizers, which could result in them misrecognizing OOV words as in-vocabulary words. To solve this issue, we propose a semi-automatic corpus-generation technique that can help to lower the number of OOV words in order to improve ASR performance.

Semi-automatic creation of the language model

In order to create an LM that considers the expression forms used in the scenario of the conversation, the corpus used in training the LM should be designed properly. In the conversations used in the experiments described in this paper, the answer uttered by R2 is considered a reference answer. The correctness of the learners' answers are measured according to the closeness to those reference answers. However, the learner may use different content in their utterances, since the main focus is the expression form. In order to consider more variety of contents in the reference answers for use as a corpus to train the LM, we suggest using WordNet.

The corpus is designed by simply collecting the reference answers into a text file. Each answer consists of the main structure of the expression form and changeable content within the form. For example, answering the question "What did you like to play in your childhood?", the reference answer used by R2 is "I liked to play soccer". The main structure of the answer consists of the pronoun "I", the verb "like" in the past tense, and the type of game chosen in this answer "soccer". The first part of the sentence "I liked to play" is not expected to change, since we are dealing with low-proficiency learners, but the last part used can (or should) be changed. WordNet is used in this case to generate sentences (in this specific example) that start with "I liked to play", while for the last part ("soccer") other possibilities may be used (e.g., "basketball", "volleyball", "tennis").

We have conducted several experiments using WordNet to find similar words that could be appropriate to generate more possible reference answers. WordNet was helpful to expand the possibilities of words used in a sentence, but it lacks the ability to provide the

proper verb conjugation. To solve this issue, the python module pattern.en from Computational Linguistics and Psycholinguistics (CLiPS) was used to ensure the proper verb form according to the grammar used. The module has a lexicon of 8,500 common English verbs and their conjugated forms.

In addition to expanding the sentences, we found that incorporating answers similar to “I don’t know” into the corpus was helpful. Another helpful addition to the corpus was the inclusion of a text of readings in English for middle school and another text of training dialogues for Japanese high school students learning English. Both included more than 5,000 vocabulary items that the Japanese learner of English might be familiar with from middle and high school.

Experiment

We conducted an experiment using the JIT-RALL system with four learners over a period of three weeks. The same setup of the experiments explained in Chapter 4 was used. We collected 358 utterances in order to perform an offline recognition using ASR. The audio files of the utterances were modified by removing all of the fillers in order to find the influence of their effect on the performance of the ASR.

We use Kaldi as the ASR engine. The AM is trained on the ERJ corpus, which is the English Speech Database Read by Japanese Students. The language model is trained using the semi-automatically generated corpus that was explained earlier.

Automatic grammar classification

An important part of the JIT-RALL system is the corrective feedback mechanism provided by R1. This feedback is controlled manually through the Wizard of Oz method. Repeating the dialogue between R1 and R2 to teach the structure of the expression form that the learner is expected to use is an implicit method of corrective feedback. In order to achieve an automatic way of providing the learner with corrective feedback, the accuracy of the ASR should be sufficiently high to detect a mistake in the learner's answer, which is currently not the case. However, exact recognition may not be necessary in our case, since the focus of the system is to convey grammatical knowledge. Accordingly, simply recognizing some parts of the utterance to detect the structure of the expression form used by the learner should suffice. We applied three mechanisms to automatically classify the learner's answers into correct answers or erroneous answers.

One way to accomplish automatic grammar classification of the learners' answers is to design a comparison structure from the reference answers uttered by R2. The main grammatical parts of the reference answer is tagged to indicate the expected content of each part and the expected order. For example, the reference answer to the question "What websites have you viewed recently?" is "I have recently viewed YouTube to watch music videos". The main structure of this answer contains 1) any number of words that may contain the pronoun "I", 2) the present tense of the verb "to have", 3) the past participle of a verb (any verb), 4) any number of words after that. The answer of the learner is checked for whether it contains these four parts and whether these parts are in order. A more detailed explanation of this mechanism can be found in the Appendix.

The second mechanism to classify the learner's answer into correct or erroneous answer is the BLEU score. BLEU is the bilingual evaluation understudy (BLEU) (Papineni et al., 2002) score. BLEU is a popular index for evaluating the quality of machine translation. It is given by the equation:

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

where p_n is the precision of n-grams in a learner utterance determined through comparison with the reference sentences, N is usually set at 4 because we do not expect a similarity of more than 4-gram to be found in a conversation, since there are no long sentences, and BP is a brevity penalty as a coefficient for correction. The answer of R2 is used as the reference sentence.

The third mechanism is the edit distance, which is the count of the minimum number of operations required to transform the learner's answer into the reference answer. An arbitrary threshold is used in each case (points calculated using the first mechanism, BLEU score, and Edit distance result) to classify the learner's answer into correct or erroneous answer.

Results

We have run ASR to recognize the utterances of the learner in an offline manner using two different LMs. The first LM (LM1) was trained using the transcription of the learners' utterances as the corpus to find the best performance the ASR could achieve. The

second LM (LM2) was trained using the semi-automatically generated corpus described earlier in this chapter, which is a real-life case of an LM.

Tables 6.1, 6.2, and 6.3 show the number of utterances using the CoU measure, explained in Chapter 5, compared to the three automatic grammar classification mechanisms explained in the previous section. The lower the precision of the classifier is, the less corrective feedback is provided to the learner. In other words, some of the erroneous utterances of the learner will not be corrected but rather will be considered correct, and thus the conversation will move on. This could cause a loss of learning opportunities for the learner. On the other hand, the lower the recall value of the classifier is, the more corrective feedback is provided to the learner, even in cases of correct utterances. This could cause the learner to become confused and discouraged. By looking at both Accuracy and F1 values, the first mechanism of automatic grammar classification showed the best result using LM1 and LM2.

Table 6.1: Number of utterances using the CoU measure, explained in Chapter 5, relative to the automatic grammar classification (AGC), identified as the first mechanism in the previous section.

		AGC for LM1				AGC for LM2	
		Correct	Erroneous			Correct	Erroneous
CoU	Correct	51	11	CoU	Correct	46	16
	Erroneous	23	214		Erroneous	18	219

Accuracy	88.63%
Precision	68.92%
Recall	82.26%
F1	75.00%

Accuracy	88.63%
Precision	71.88%
Recall	74.19%
F1	73.02%

Table 6.2: Number of utterances using the CoU measure, explained in Chapter 5, relative to the BLEU score classifier.

		BLEU for LM1				BLEU for LM2	
		Correct	Erroneous			Correct	Erroneous
CoU	Correct	26	36	CoU	Correct	23	39
	Erroneous	9	228		Erroneous	6	231

Accuracy	84.95%
Precision	74.29%
Recall	41.94%
F1	53.61%

Accuracy	84.95%
Precision	79.31%
Recall	37.10%
F1	50.55%

Table 6.3: Number of utterances using the CoU measure, explained in chapter 5, relative to the edit distance classifier.

		Edit Distance for LM1				Edit Distance for LM2	
		Correct	Erroneous			Correct	Erroneous
CoU	Correct	39	23	CoU	Correct	37	25
	Erroneous	17	220		Erroneous	10	227

Accuracy	86.62%
Precision	69.64%
Recall	62.90%
F1	66.10%

Accuracy	88.29%
Precision	78.72%
Recall	59.68%
F1	67.89%

Discussion

The different mechanisms of automatic grammar classification presented in this chapter showed promising results. Even with the low accuracy of the ASR, which is 46.77 (WER = 53.23%) for LM1 and 28.61% (WER = 71.39%) for LM2, however, the best F1 value of the classifiers reached 75%. It could be that the ASR engine is mis-recognizing many of

the “non-important words” unrelated to the expression form expected from the learner while correctly recognizing the “important” words from a grammatical perspective. The main focus of the system is to convey the grammatical knowledge of the expression form chosen for the conversation. In this case, the exact recognition of the ASR engine may not be required; however, the main parts of the expression form are needed. The automatic grammar classification mechanisms suggested could achieve promising results in this case.

Another main factor that affects the accuracy of the ASR engine is the AM. We have focused on designing a semi-automatically generated corpus to enhance the accuracy of the ASR engine; however, some improvement could be made for the audio data. As discussed in Chapter 5, the accuracy of the ASR engine could be enhanced by about 20% by simply removing the fillers from the audio files. Encouraging the learner to speak with more confidence, by asking him/her to repeat the answer again instead of just giving a spontaneous answer, could help to obtain utterances with fewer fillers.

Chapter 7

7. Conclusion and future work

The main contribution of this work was the design of a joining-in-type robot-assisted language learning (JIT-RALL) system. Two robots were used to conduct a goal-oriented conversation with a human learner in L2. The system uses implicit learning as the main learning style to convey implicit grammatical knowledge of how to use a specific expression form. A mix of tutoring and peer learning is implemented in the framework of a three-party conversation.

A series of experiments using the JIT-RALL system was presented. These experiments were conducted to measure the effect of implicit learning through the system conveying expression forms to human learners. Another contribution of the experiments was measuring the effect of repetitive queries on implicit learning. We could also measure the human learner's ability to retain what was learned implicitly.

We conducted preliminary analyses on introducing an ASR in the JIT-RALL system. We could achieve promising results in automatically classifying the recognized utterances of the learners based on their grammatical correctness. This classification of utterances was based on a single factor. Using a more sophisticated classifier that incorporates

multiple factors could result in better classification accuracy. This achievement could enable us to develop an automatic corrective feedback mechanism for use during a conversation that would replace the current Wizard of Oz method.

For the JIT-RALL system to provide automatic corrective feedback, the ASR performance should be handled carefully, since the output of recognition would affect the level of feedback provided to the learner. The system's educational outcome of regulating the learners' utterances will be applied to raising the level of the utterances' predictability so that the ASR engine could achieve yet higher performance.

The fillers in the utterances of the learners were shown to affect the accuracy of the ASR engine. We plan to consider this issue when building the ASR engine. For example, the structure of the conversation could help the learners to produce fewer fillers in their utterances by asking them to repeat their answers a couple of times in order to increase their confidence. Considering the fillers when building the LM could be another way to increase the accuracy of the ASR engine.

The automatic grammatical classification that was built on top of the ASR could be used by the robots to conduct conversations in a fully autonomous manner. After reaching that capability, the scenarios used by the system would be more flexible and thus include a wider range of topics, at the same time providing a wider range of linguistic features for the learner's L2 practice.

Acknowledgments

I would like to express my deepest gratitude to my advisor Professor Seiichi Yamamoto for his encouragement, support and continual guidance throughout my years at Doshisha University. His wisdom, insights, and thoughtful comments, and all of his efforts, have definitely improved the quality of my knowledge and my work. He always provided me with intelligent guidance to investigate and explore this field with a view to my future directions. Without my dear advisor, I would not have been inspired by this study, nor enjoy it and be excited about this research.

I must thank Professor Tsuneo Kato for his support and advice, which helped me proceed with the various stages of this research. My skills and my work were improved because of him. I was inspired by his activity in working hard and producing results effectively.

I also thank Dr. Ichiro Umata and Prof. Graham Wilcock for their comments and encouragement, which greatly helped me in this research.

Last but not least, I would like to express my deep gratitude to my family, my wife, and my children for their enormous support and encouragement.

Bibliography

[1] Alemi, M., Meghdari, A., & Ghazisaedy, M. (2014). Employing humanoid robots for teaching English language in Iranian junior high-schools. International Journal of Humanoid Robotics, 11(03), 1450022.

[2] Anderson, J. N., Davidson, N., Morton, H., & Jack, M. A. (2008). Language learning with interactive virtual agent scenarios and speech recognition: Lessons learned. Computer Animation and Virtual Worlds, 19(5), 605-619.

[3] Baxter, P., Kennedy, J., Ashurst, E., & Belpaeme, T. (2016). The Effect of Repeating Tasks on Performance Levels in Mediated Child-Robot Interactions. Proc. of Robots 4 Education Workshop 2016.

[4] Belpaeme, T., Kennedy, J., Baxter, P., Vogt, P., Krahmer, E. E., Kopp, S., Bergmann, K., Leseman, P., Küntay, A. C., Göksun, T., Pandey, A. K., Gelin, R., Koudelkova, P., & Deblieck, T. (2015). L2TOR-Second language tutoring using social robots. Proc. of Int. Workshop on Educational Robots.

[5] Bibauw, S., François, T., & Desmet, P. (2019). Discussing with a computer to practice a foreign language: research synthesis and conceptual framework of dialogue-based CALL. Computer Assisted Language Learning, 1-51.

[6] Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational researcher, 13(6), 4-16.

[7] Brennan, S. E. & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(6), 1482. https://doi.org/10.1037/0278-7393.22.6.1482

[8] Dalby, J., & Kewley-Port, D. (1999). Explicit pronunciation training using automatic speech recognition technology. CALICO journal, 425-445.

[9] Ellis, N. C., & Bogart, P. S. (2007). Speech and Language Technology in Education: the perspective from SLA research and practice. In Workshop on Speech and Language Technology in Education.

[10] Fandrianto, A. & Eskenazi, M. (2012). Prosodic Entrainment in an Information-driven Dialog System. In INTERSPEECH, p. 342-345.

[11] Hjalmarsson, A., Wik, P., & Brusk, J. (2007). Dealing with DEAL: a dialogue system for conversation training. In Proceedings of SIGDIAL (pp. 132-135).

[12] Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. Human–Computer Interaction, 19(1-2), 61-84.

[13] Kennedy, J., Baxter, P., Senft, E., & Belpaeme, T. (March 2016). Social robot tutoring for child second-language learning. Proc. of Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference, 231-238.

[14] Khalifa, A., Kato, T., & Yamamoto, S. (2017). Measuring Effect of Repetitive Queries and Implicit Learning with Joining-in-type Robot Assisted Language Learning System. In SLaTE (pp. 13-17).

[15] Khalifa, A., Kato, T., & Yamamoto, S. (2018, September). The Retention Effect of Learning Grammatical Patterns Implicitly Using Joining-in-Type Robot-Assisted Language-Learning System. In International Conference on Text, Speech, and Dialogue (pp. 492-499). Springer, Cham.

[16] Khalifa, A., Kato, T., & Yamamoto, S. (2019). Learning Effect of Implicit Learning in Joining-in-type Robot-assisted Language Learning System. International Journal of Emerging Technologies in Learning, 14(2).

[17] Knill, K. M., Gales, M. J. F., Manakul, P. P., & Caines, A. P. (2019, April). Automatic Grammatical Error Detection of Non-native Spoken Learner English. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8127-8131). IEEE.

[18] Knill, K. M., Gales, M. J. F., Manakul, P. P., & Caines, A. P. (2019, May). Automatic Grammatical Error Detection of Non-native Spoken Learner English. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8127-8131). IEEE.

[19] Lee, K., Kweon, S. O., Lee, S., Noh, H., & Lee, G. G. (2014). POSTECH immersive English study (POMY): Dialog-based language learning game. IEICE TRANSACTIONS on Information and Systems, 97(7), 1830-1841.

[20] Lee, S., Noh, H., Lee, J., Lee, K., Lee, G. G., Sagong, S., & Kim, M. (2011). On the effectiveness of robot-assisted language learning. ReCALL, 23(1), 25-58.

[21] Leyzberg, D., Spaulding, S., Toneva, M., & Scassellati, B. (2012). The physical presence of a robot tutor increases cognitive learning gains. In Proceedings of the annual meeting of the cognitive science society (Vol. 34, No. 34).

[22] Liu, M., Moore, Z., Graham, L., & Lee, S. (2002). A look at the research on computer-based technology use in second language learning: A review of the literature from 1990–2000. Journal of research on technology in education, 34(3), 250-273.

[23] Mizukami, M., Yoshino, K., Neubig, G., Traum, D., & Nakamura, S.

(September 2016). Analyzing the Effect of Entrainment on Dialogue Acts. In 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue SIGDIAL, p. 310-318. https://doi.org/10.18653/v1/W16-3640

[24] Neri, A., Cucchiarini, C., & Strik, H. (2001). Effective feedback on L2 pronunciation in ASR-based CALL. Proc. of the workshop on Computer Assisted Language Learning, Artificial Intelligence in Education Conference. pp 40-48.

[25] Niederhoffer, K. G. & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. Journal of Language and Social Psychology, 21(4), 337-360. https://doi.org/10.1177/026192702237953

[26] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics.

[27] Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and brain sciences, 27(2), 169-190.

[28] Ping, T. T. (2008). Automatic Speech Recognition for Non-Native Speakers (Doctoral dissertation, Université Joseph-Fourier-Grenoble I).

[29] Raux, A., & Eskenazi, M. (2004). Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges. In InSTIL/ICALL Symposium 2004.

[30] Reber, A. S. (1967). Implicit learning of artificial grammars. Journal of verbal learning and verbal behavior, 6(6), 855-863.

[31] Reitter, D. & Moore, J. D. (2007). Predicting success in dialogue. Proc. of Annual Meeting of ACL 2007, p. 808-815

[32] Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., & Zue, V. (1998).

Galaxy-II: A reference architecture for conversational system development. In Fifth International Conference on Spoken Language Processing.

[33] Seneff, S., Wang, C., & Chao, C. Y. (2007, April). Spoken dialogue systems for language learning. In Proceedings of human language technologies: The annual conference of the north American chapter of the association for computational linguistics: Demonstrations (pp. 13-14). Association for Computational Linguistics.

[34] Seneff, S., Wang, C., & Zhang, J. (2004). Spoken conversational interaction for language learning. In InSTIL/ICALL Symposium 2004.

[35] Wilcock, G. & Yamamoto, S. (2015). Towards computer-assisted language learning with robots, Wikipedia and CogInfoCom. In Cognitive Infocommunications (CogInfoCom), 2015 6th IEEE International Conference on pp. 115-119. IEEE. https://doi.org/10.1109/CogInfoCom.2015.7390575

[36] Yamamoto, S., Taguchi, K., Ijuin, K., Umata, I., & Nishida, M. (2015). Multimodal corpus of multiparty conversations in L1 and L2 languages and findings obtained from it. Language Resources and Evaluation, 49(4), 857-882.

Appendix: Automatic Grammar Classification

The mechanism of automatic grammar classification (AGC) is here suggested for use in evaluating the answers of human learners by the JIT-RALL system in an automatic manner. It would also be helpful for supporting the automatic corrective feedback feature, which is considered a major part of the system.

AGC is appropriate for use with a manually designed conversation scenario because it requires manual tagging of the expected answer by the learner. Each answer by R2 follows a specific grammatical expression form among other expression forms chosen for every conversation. The human learner is expected to learn it and then answer according to that expression form. Evaluating the answer of the learner using AGC would be more accurate than the other methods suggested in this paper (i.e., BLEU score and Edit Distance).

In every answer by R2, there are some words that are considered more important than others for the evaluation, from the grammatical perspective. There are some fixed parts, where the exact words are expected to be used, since the system assumes low-proficiency learners of English. Some parts are changeable, like verbs, where the right tense is expected to be used but not the exact verb. In addition to the existence of the important parts in the answer, the order of the parts is also an important factor. The

learner may add more words that do not affect the evaluation from a grammatical perspective, and these should not affect the results of the evaluation.

Table A.1 shows examples of tagging reference answers uttered by R2. The first example represents a past tense expression form. The tagged form consists of three parts:

1. The first part, tagged by empty square brackets, could contain any number of words such as “I” or “I think I”. This part has to contain at least one word, but no specific word is checked.
2. The second part is tagged by square brackets containing the letter v and the number 4 in angular brackets, which means that it has to contain a verb in the past tense, which could be “liked” or “played” or any other verb. The letter v represents a verb, and the number 4 represents a verb in the past tense. Other possible numbers are:
    1) For the basic form of the verb (e.g., drive)
    2) For the third-person form of the verb (e.g., drives)
    3) For the continuous form of the verb (e.g., driving)
    4) For the past tense form of the verb (e.g., drove)
    5) For the past participle form of the verb (e.g., driven)

These numbers are arbitrarily chosen in a python code written specifically for the AGC evaluation, and they match the verb conjugation found in the python module pattern.en from Computational Linguistics and Psycholinguistics (CLiPS). WordNet is also used to check whether the word is a verb.

3. The third part is the same as the first part.

The second example in Table A.1 represents a present perfect simple expression form. The same description can be given in this example as in the first example, with the difference that the word “have” is a fixed mandatory word that must exist in the answer. Another difference is that the verb form expected is the participle form. The fourth example in the table shows a choice for the first part of the tagged form between “yes” and “no”.

Table A.1: Examples of expression forms and their tagged forms.

#	Question by R1	Answer by R2	Tagged Form
1	What game did you like to play in childhood?	I liked to play soccer.	[] [<v4>] []
2	What kind of videos have you watched recently on YouTube?	I have watched music videos and live videos.	[] [have] [<v5>] []
3	What do you think makes him go abroad?	New exciting experiences make him go abroad.	[] [makes] [him] []
4	Are you going to buy any electronic devices this year?	Yes, I’m just going to buy a smart watch.	[yes/no] [] [going] []

AGC could perform better than the BLEU score or the Edit Distance measures used in this paper because it focuses on the grammatical parts of the sentence. The other two mechanisms make a fixed comparison between the reference answer and the human learner’s answer, which gives no consideration to the changeable parts in the answer.

On the other hand, AGC has to be prepared manually, which may not be possible when the scenario is dynamic. The JIT-RALL system is expected to have a more flexible topic-changing mechanism in the future, and thus a different evaluation mechanism should be designed for handling that condition.