Korean Journal of English Language and Linguistics, Vol 21, April 2021, pp. 375-391
DOI: 10.15738/kjell.21..202104.375

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474
http://journal.kasell.or.kr

Exploring the Use of An Artificial Intelligence Chatbot as Second Language Conversation Partners*

Dongkwang Shin (Gwangju National University of Education) **Heyoung Kim** (Chung-Ang University) **Jang Ho Lee** (Chung-Ang University) **Hyejin Yang** (Chung-Ang University)

Received: March 23, 2021
Revised: April 20, 2021
Accepted: April 25, 2021

Dongkwang Shin (1st author)
Professor, Gwangju National Univ. of Education
sdhera@gamail.com

Heyoung Kim (corresponding author)
Professor, Chung-Ang Univ.
englishnet@cau.ac.kr

Jang Ho Lee (co-author)
Professor, Chung-Ang Univ.
jangholee@cau.ac.kr

Hyejin Yang (co-author)
Researcher, Chung-Ang Univ.
hjyang1112@gmail.com

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A03037255).

ABSTRACT

Shin, Dongkwang, Heyoung Kim, Jang Ho Lee and Hyejin Yang. 2021. Exploring the use of an artificial intelligence chatbot as second language conversation partners. *Korean Journal of English Language and Linguistics* 21, 375-391.

This study investigated the appropriateness of using artificially intelligent chatbots as conversation partners for second language (L2) learners. 27 Korean high school and 26 college students had a task-oriented conversation with a text-based chatbot, Mitsuku, for 20 minutes. Chat log data were collected and analyzed quantitatively and qualitatively in terms of the quantity of students' utterances and their vocabulary levels, along with the degree of conversation task success between the chatbot and its users. Both groups finished their tasks, successfully developing conversations with the chatbot and producing double the expected minimum quantity of utterances, although their performances varied individually. Mitsuku's vocabulary was deemed appropriate for L2 learners' proficiency. The college students used conversational strategies more appropriately than their high school counterparts. Nevertheless, a sentiment analysis showed that the high school students enjoyed talking with Mitsuku to a greater extent than the college students. These results suggest that the chatbot offers L2 learners substantial opportunities as a conversation partner.

KEYWORDS

Artificial Intelligence (AI), chatbot, Mitsuku, conversation task, vocabulary level, task success rate, sentiment analysis, Orange 3.28.0

1. Introduction

Artificial Intelligence (AI) chatbots have been in development since the 1960s and are now utilized in various fields, replacing human beings. AI chatbots are integrated into a variety of devices and platforms including Amazon's Echo, Google Home, Internet chatbot sites (e.g., Pandorabot and Cleverbot), Social Networking Services (e.g., Kakao Talkbot, XiaoIce), and Intelligent Personal Assistants (IPAs), such as Apple's Siri or Samsung's Bixby. AI chatbots have been applied for a range of purposes, including 24-hour customer consultation services, purchasing airline tickets, writing newspaper articles, and predicting the stock market or sports results.

Some researchers have studied chatbots as first language communication partners (e.g., Batacharia Levy, Catizone, Krotov and Wilks 1999, Colby 1999, Wallace 2009, Weizenbaum 1966), but just a few have investigated their roles in second language (L2) education (e.g., Coniam 2014, Kwon, Lee, Kim and Lee 2015). Conversing with an AI chatbot in their target language could highly motivate L2 learners and provide them with an effective practice opportunity. The appropriate usage of L1 chatbots may assist L2 learners in language acquisition (Abu-Shawar 2017, Lu, Chiou, Day, Ong and Hsu 2006). To enable good conversations between English speaking chatbots and L2 learners, the chatbot must be able to provide "comprehensible input" (Krashen 1980, 1981) relative to the learners' L2 proficiency in the form of realistic exchanges. Therefore, effective communication is determined by 1) whether the chatbot can adjust its language use for compatibility with L2 learners and 2) whether L2 learners can successfully negotiate the meanings by interacting with it.

Despite the fact that chatbots offer many advantages in the field of L2 education, there have been a few cases where a chatbot program was integrated into an L2 curriculum and no further studies traced the efficacy of the instructional design or the suitability of the learning conditions. In particular, there is a paucity of studies that compare learner groups with differing L2 proficiencies and focus on how well they are able to exchange meanings with an L1 chatbot. For successful conversation development, it is important to design second-language learning tasks that foster meaning negotiation and suit L2 learners' interests and proficiency levels; understanding how these tasks might best be designed and implemented, then, will contribute to successful conversation development (Lee, Yang, Shin and Kim 2020).

Mitsuku is not a chatbot specifically developed for language learning. The main purpose of this study is to investigate whether a free accessible chatbot such as Mitsuku is an appropriate tool for language learning for Korean L2 learners who do not reach the proficiency level of native English speakers in terms of the natural linkage of the interaction between the chatbot and the learner, the appropriate level of the input language uttered by the chatbot, and the affective domain of the learner during the conversation with the chatbot. To this end, 53 English learners in Korea, with different degrees of proficiency in English, engaged in individual chats with "Mitsuku," a widely recognized text-based AI chatbot (accessible at https://www.pandorabots.com/mitsuku/).

Participants were allotted a maximum of 20 minutes to complete seven conversation tasks designed by the researchers. To identify Mitsuku's text characteristics, including the sophistication of her

vocabulary and her responses to L2 users' utterances, the chat dialogs generated by two groups of students while performing L2 conversation tasks were analyzed extensively in terms of the total amount of utterances, the number of conversation turns, the vocabulary level, emotional attitude towards conversations with the chatbot, and the task success rates. The qualitative data were also provided for both further speculation and the triangulation of the data analysis. The following research questions were used to investigate the abovementioned critical issues in implementing chatbots in the L2 classroom:

Research Question 1: How would students with different levels of L2 proficiency develop their conversations with the chatbot?

Research Question 2: Could the chatbot's language be suitable for students with different levels of L2 proficiency?

Research Question 3: What is the degree of satisfaction of students for chatbot-based tasks through a sentiment analysis?

2. Literature Review

2.1 Status of Chatbot Development

From early chatbots like Eliza and Parry in the 1960s and 1970s to the task-executing dialog systems of the 2000s, anthropomorphic communication systems have continued to advance (Wallace 2009, Wang and Petrina 2013, Weizenbaum 1966). Over the course of the 2010s, personal assistant chatbots with sociolinguistic capabilities, such as Siri and XiaoIce, have become increasingly common (Fryer, Coniam, Carpenter and Lăpuşneanu 2020).

The first ever chatbot, Eliza, developed in 1966 by Joseph Weizenbaum of MIT, responded to text-based input. Its responses were based on its analysis of dialog expression patterns rather than actual meaning, resulting in a limited response capacity (Weizenbaum 1966). Since then, task-completion systems have been developed (Dahl et al. 1994) that are unlike traditional dialog systems which simply parrot preprogrammed data. Task-completion systems are designed to limit conversations to specific areas. For example, in the DARPA Communicator Programs, chatbots require data relevant only to helping customers make airline reservations. To extract key meanings from customers' utterances, Spoken Language Understanding (SLU) systems convert voice to text via an Automatic Speech Recognizer (ASR). A Dialog Manager then categorizes the utterance and analyzes its meaning and intention. Once the user's needs are identified, the system searches its database for an appropriate response. Finally, the Natural Language Generator produces the response text, and the Text to Speech converter converts it to an utterance, which is received by the customer. A linear structure was initially adopted, but recent programs have evolved parallel structures that produce optimized responses by

applying multiple variables (Fryer et al. 2020).

Artificial Linguistic Internet Computer Entity (A.L.I.C.E), the natural language chatbot famously developed by Richard Wallace in 1995, is a model for Mitsuku. A.L.I.C.E uses Artificial Intelligence Markup Language (AIML) to analyze the dialog patterns of input language data and converts such data into anthropomorphic speech utterances by locating heuristic pattern matching rules (Wallace 2009). A.L.I.C.E has been continuously updated since its inception and is also free to use.

Since Apple launched Siri in 2011, various IPAs have been developed, including Microsoft's Cortana, Google Assistant, and Amazon's Alexa. IPAs integrate information collected from sensors for items such as location, time, movement, touch, gestures, and gaze. These data, combined with information obtained by accessing users' music, movies, calendars, e-mails, and personal profiles, enable IPAs to offer users various services, such as sending messages, scheduling meetings, and finding venues. In addition, IPAs are designed to browse the web for alternatives when they cannot immediately meet users' needs. They also provide reactive services such as information search and restaurant reservations as well as proactive services such as schedule reminders or recommendations for specific items. IPAs have evolved simultaneously with mobile devices, personal computers, and smart home devices.

Mitsuku exemplifies this evolution. Among chatbots, her utterances are considered the most anthropomorphic, and she won the prestigious Loebner Prize each year from 2016 to 2019. Therefore, Mitsuku was selected as the target chatbot of this study.

2.2 Application of Chatbots in L2 Learning

The Computer-Assisted Language Learning (CALL) framework has been applied to L2 learning as a model for distance learning and for teaching language by assisting teachers (Compton 2009). CALL offers lessons and activities designed to enhance learners' vocabulary, grammar, and writing skills and provides immediate feedback to language learners. Although CALL facilitates students' independent learning, L2 learners still need to enhance their communication skills through natural, or naturalistic, conversations. Advanced chatbots are expected to meet this need by serving as conversation partners for L2 learners. Conversing with chatbots has many advantages for language learners, such as improved learner autonomy and reduced learner anxiety. Chatbots are convenient as users can converse with them regardless of the time and location. Additionally, they provide records of conversations for future review (Lu et al. 2006).

Lu et al. (2006) suggested that learners register chatbots as friends on social networking sites and engage in regular conversations with them to improve their L2 abilities. XiaoIce (meaning Little Ice), the Chinese-speaking chatbot released by Microsoft in 2014, was modeled on the chatbot proposed by Lu et al. (2006). XiaoIce is a social chatbot and was launched through China's leading instant messaging applications such as WeChat and Weibo and has already been used by millions of people.

Among recent studies on chatbots, Abu-Shawar (2017) developed an extended model for dialogs on various topics by integrating the Corpus of Spoken Afrikaans into AIML. An Automatic AIML Generator, which converts data into a format that A.L.I.C.E can understand, has also been developed. In Abu-Shawar's study,

1,256 South Africans participated in a conversation experiment using A.L.I.C.E. While 17% of the participants responded positively to A.L.I.C.E as their conversation partner, 24% responded negatively and 59% demonstrated a neutral attitude. Topics of conversation included “exams and English study (11.39%)” and “love (13%).” Abu-Shawar argued that even a chatbot operating at the relatively simplistic level of keyword matching can communicate with users about diverse topics.

As a more recent study, Alm and Nkomo (2020) analyzed the user reviews of online L2 learning applications that include chatbot functions-Duolingo, Eggbun, Memrise, and Mondly-to measure the degree of the users’ satisfaction of these applications. For this analysis, the sentiment analyzer of Microsoft Azure Cognitive Services was used. As a result, it turned out that the users showed some levels of willingness to participate in conversations with a chatbot, and the reason was mainly attributed to their curiosity about chatbots.

3. Data and Methodology

This section first introduces participants, two student groups and a chatbot, and explains seven chatbot-based conversation tasks. Next, data collection and analysis methods are introduced with the concordance programs employed in this study.

3.1 Participants

The participants were divided into upper and lower groups as shown in Table 1. Group one consisted of college juniors (n = 27) from a college in Gwangju, and group two consisted of high school sophomores (n = 26) from a high school in Seoul, Korea.

Table 1. Participants’ Backgrounds

Group (N = 53)	Age	L2 Level
College Students (n = 27)	21-24	Upper
High School Students (n = 26)	17	Lower

The college students were from a high-ranked university, and the high school students were attending a general high school in South Korea. Even though participants' level of English proficiency, especially writing skill, was not measured, there might be a considerable difference in English proficiency between the two groups. Besides, the focus of the present study was to see if the English input by Mitsuku is appropriate for the level of Korean college students or high school students, thus the English proficiency of individual learners as a research variable was not in the scope of this study. Both groups were taking English writing courses at their respective schools at the time of the study. The experiment was performed with the consent of the participants, and the names of participants mentioned in this paper are pseudonyms.

3.2 Target Chatbot

Mitsuku, characterized as an English speaking 18-year-old woman living in Leeds, England, was developed by Steve Worswick in 2005 and has since been constantly upgraded. She recognizes students' utterances, extracts keywords from their conversations, and then analyzes the attributes of these keywords to perform logical judgments. Her ability to search the web allows her to access information outside of the database stored in AIML. This contributes to her ability to handle multiple questions and tasks simultaneously (AiDreams 2013).

3.3 Tasks

All participants were given the same seven specific tasks (comprising 12 questions) to complete within 20 minutes. At both schools, the experiment was conducted in a computer lab during the session that would normally be scheduled for the participants' English writing course. The tasks were primarily adapted from the Oral Proficiency Interview from the International English Language Testing System (IELTS 2018), one of the most widely used standardized measures of proficiency in spoken English. The seven tasks were sequenced according to the flow of general conversation, that is, the opening content, followed by the main content (exchanging personal questions to ask for general information) and then the closing content. The instructions and the expected minimum number of conversation turns based on sub-questions required for each task are summarized in Table 2.

Table 2. Instructions for Each Task and the Expected Number of Minimum Conversation Turns

#	Type of Content	Direction	Number of Minimum Conversation Turns
1	Opening	Greet Mitsuku.	1
2		Introduce yourself to Mitsuku.	1
3	Main (3-6)	Find out where Mitsuku lives.	1
4		Find out about Mitsuku's job and what Mitsuku can do to help people.	2
5		Find out what Mitsuku does in her free time (e.g., reading, watching TV, sports, art, music, food). Ask at least two follow-up questions about Mitsuku's hobbies.	3
6		Find out what happened 10 years ago today, and then select an event to learn more about. (e.g., If you ask what happened on September 6th 10 years ago, you may find information about several events on that day including, 'the first cricket Test Match in England.' If you want to find out more about that, ask what cricket is, whether Mitsuku likes cricket, or who Mitsuku's favorite cricket player is.)	3-4
7	Closing	Finish conversations.	1

3.4 Data Analysis

This study's primary data comprised the chat logs of the conversations between the participants and the chatbot, Mitsuku. The participant conversation data were all collected and coded by the two proficiency groups. They were first analyzed descriptively by counting the number of words produced by both Mitsuku and each

participant, and the completion rate for each task as well as the number of turns the participants took to complete each task were counted. In addition, the BNC-COCA 25000, Range Program, and Orange 3.28.0 were employed to analyze the data. To analyze the quantity of Mitsuku and the participants' utterances as well as the level of vocabulary they employed, BNC-COCA 25,000 (Nation 2012) was used, loaded on the RANGE program (Heatley, Nation and Coxhead 2002). BNC-COCA 25,000 is the most representative vocabulary list available and is based on the lexical information of the British National Corpus (BNC) and the US Corpus of Contemporary American English (COCA) (Shin 2014). Given that the participants were non-native speakers of English, the analysis was only applied to words at the level of 4000.

The degree of satisfaction on the chatbot-based tasks was measured through sentiment analysis. To this end, Orange 3.28.0 (Demsar et al. 2021), a free software for data mining, was used to measure the emotional state of the students during conversation with Mitsuku. The sentiment analysis on Orange 3.28.0 can analyze positive, negative, and neutral emotional states based on the VADER sentiment analysis algorithm-sentiment dictionary and rule-based sentiment analysis-developed by Hutto and Gilbert (2014). A compound score close to 1 means positive, close to 0 means neutral, and close to -1 means negative. In simpler words, positive scores about conversations with the chatbot could imply that the learners enjoyed engaging in conversations with the chatbot.

4. Results and Discussion

The findings are reported by the sequence of research questions: namely, conversation development, vocabulary level, task success, and satisfaction with conversation tasks.

4.1 Research Question 1: How would students with different levels of L2 proficiency develop their conversations with the chatbot?

First, we retrieved descriptive statistics of the total number of words used and turns taken by the students and Mitsuku that was relevant to understanding how the participants developed their conversations. Table 3 presents these descriptive data.

Table 3. Descriptive Statistics on the Number of Words and Conversation Turns During Chatbot Tasks

	From	To	*Mean* (*SD*)	Min/Max
No. of Words (Token)	College	Mitsuku	255 (99.03)	91/518
	High School		159 (96.64)	55/407
	All Students		207.85 (108.01)	55/518
	Mitsuku	College	621 (227.01)	203/1054
		High School	436 (199.98)	110/949
		All Students	530.72 (232.18)	110/1054
No. of Conversation Turns	Mitsuku-Students	College	32.185 (11.75)	14/65
		High School	24 (15.05)	7/60
		All Students	28.17 (13.24)	7/65

Overall, the average number of total tokens and conversation turns by the students was 207.85 and 28.17, respectively. This means that the students uttered about seven words per turn (specifically, college students uttered 7.97 words and high school students 6.63 words) and exchanged four turns per task, which was double the required minimum number of conversation turns. Most of the participants (n = 47) successfully completed all the tasks.

More specifically, students' utterances significantly varied in both groups (min/max = 55/518 words). In the college group, the participant who talked the most uttered 518 words, whereas the one who talked the least uttered 91 words, about five times less, to complete the same task goals. In the high school group, an even greater difference existed between the maximum (407 words) and minimum (55 words) number of words contributed. With regard to conversation turns, individual students exhibited extremely varied performances (min/max = 7/65). Only six high school students missed a question that was important for task completion. In this case, the students tried to ask a question once or twice, but the chatbot did not understand it correctly, then they gave up the task. Nevertheless, the rest (n = 47) attained task completion, making rich conversation.

Second, comparing the number of utterances exchanged between the participants and the chatbot, it is evident that the participants generally "listened more and talked less." During the 20-minute session, the average number of total tokens produced by the students was 207.85 (SD = 108.01 words) and, by Mitsuku, 530.72 (SD = 232.18 words). The average conversation share rate was 28% (207.85 words) of the total utterances (738.57 words). This result may suggest that the students spent more than two-thirds of the session time receiving input from the chatbot and one-third producing output. This is consistent with the finding of Shin's (2019) study that Mitsuku is designed to be able to answer multiple questions at the same time, so sometimes it tends to present too much input at once without considering the context.

Third, when comparing the two participant groups, the college students generally produced a greater number of words (M = 255 words) than their high school counterparts (M = 159 words). Thus, not surprisingly, college students conversed with Mitsuku significantly more than high school students. Mitsuku also produced more words in conversation with the college students (M = 621 words) than with the high school students (M = 436 words). In addition, the number of turns taken between the college students and Mitsuku was about 33% (M = 32 turns), higher than that between the high school students and Mitsuku (M = 24 turns). Thus, college students produced 7.97 words per turn while high school students produced 6.63, which does not indicate a huge difference between the groups. Therefore, the significant difference in utterance between the two groups was in the number of conversation turns.

Furthermore, as shown in Table 4, the conversation flow (opening-main-closing) between the students and Mitsuku went well, mimicking the normal development of human conversational discourse. Overall, 80% of the utterances (204 words out of 255) of the college group and 83% (132 words out of 159) of the high school group were performed during the main tasks (3-6). Mitsuku also talked more (538 words with the college group and 378 with the high school group) during the main tasks. In terms of their respective contributions, the students in both groups more actively asked questions and gave responses to Mitsuku's answers during the given tasks. The college students exchanged 26.11 turns (8.7 turns per task) and high school students 19.50 (6.5 turns per task) with Mitsuku during the main tasks, while the college group took two turns per task and the high school group only one per task during the opening and closing tasks.

Table 4. Number of Words Used and Conversation Turns Taken by Mitsuku and the Participants in Accordance with the Task Type

		From College	From High School	From Mitsuku	
		Mean (*SD*)		*Mean* (*SD*)	
	Task Type (#)	To Mitsuku		To College	To High School
No. of Words (Token)	Opening (1−2)	39 (42.975)	19 (24.280)	72 (65.426)	47 (54.664)
	Main (3−6)	204 (89.795)	132 (86.949)	538 (206.143)	378 (173.295)
	Closing (7)	12 (12.251)	8 (4.878)	11 (9.897)	11 (7.594)
		Mitsuku-Students			
		Mean (*SD*)			
	Task Type (#)	College		High school	
No. of Turns	Opening (1−2)	4.15 (3.45)		2.54 (1.43)	
	Main (3−6)	26.11 (11.18)		19.50 (13.88)	
	Closing (7)	1.93 (0.83)		1.96 (1.52)	

In their study on classroom discourse, Behnam and Pouriran (2009) showed that speakers' turn-taking presented the exchange of knowledge in discourse and active meaning negotiation, which subsequently promoted learners' progress in language learning. Thus, the high number of turns taken in the present study could constitute as evidence for Mitsuku acting as an effective conversational partner for L2 learning.

Excerpts 1 and 2 present examples of the chat logs between the two groups to show how their conversations developed according to the number of their utterances. These logs cite the conversation exchanged in Task 5, part of the main conversation, and were taken from the two students who produced the number of total tokens closest to the average for each group, in order to reflect how conversation tended to develop for each group.

Excerpt 1

Won-Ho, a High School Student

# of tokens: 53/170 (Task 5/total)

17 S I'll ask you a question again. What do you do in your free time?

18 M Once more? You can ask me anything you want. I don't really get a lot of free time as there is always someone in the internet who wants to talk to me. Any spare time I get, I use to go through my old logs and see if I can learn anything.

19 S Nice. What's your favorite sports?

20 M Thank you for that remark. football.

21 S Which football team do you like? And why do you like it?

22 M Glory! Glory Leeds United! And what? I have a lot of good experience with it.

Excerpt 2

Yu-Cheon, a College Student

# of tokens: 84/276 (Task 5/total)

23 S What is your hobby? My hobby is playing soccer. Soccer is my favorite sports.

24 M My hobby is chatting online. That sounds like a fun way to spend your free time. How long have you been interested in playing soccer? I like watching soccer. It gets really exciting sometimes.

25 S I heard that you live in Leeds, I know the football club, Leeds United.

26 M Who told you that? I think someone is spreading rumours about me.

27 S I have played soccer since when I was young. You said that you like watching soccer, then what is your favorite team or

23	S	Tell me about your good experience with Glory Leeds United.
24	M	with Glory Leeds United = Glory Leeds United came with my personality.
25	S	Do you like reading? Please recommend your favorite book to me.
26	M	I enjoy scanning books to gain knowledge. Wow you are very polite. I don't have any special my favorite book to you to recommend.
27	S	I like Leeds United because I live in Leeds. Do you like soccer player? I think you like Leeds United because you live in Leeds,
28	M	Do you play soccer since when you were young professionally or just for fun? My favorite show is Star Trek Voyager. Me? Of course I do. I love to watch Leeds United play football. I am in Mousebreaker's home in Leeds. Where are you?
29	S	I am playing soccer just for fun. Why do you like Leeds United?
30	M	In England they call it football. I like Leeds United. Because I live in Leeds. They are my local side.
31	S	Oh, I like my local football team, too.

* *Note*. S: Student, M: Mitsuku

Excerpts 1 and 2 confirmed that two students' conversations with the chatbot went smoothly without any breakdown or misunderstanding of each other, and both of them exchanged five contributions, the ideal number. Still, there was some difference between the two students. In the case of Yu-Cheon, as can be seen in contribution 27 in Excerpt 2, he informed the chatbot of his interests in soccer and also made a reference to Leeds United, which was associated with where Mitsuku lived. This move made by Yu-Cheon led Mitsuku to ask a follow-up question in contribution 28 (*Do you play soccer since when you were young professionally or just for fun?*), which in turn contributed to further conversation development on the topic (e.g., hobbies and soccer). This contrasts somewhat with Won-Ho's conversation with Mitsuku, in which Won-Ho mostly took on the role of the interviewer (presumably with the aim of completing the tasks), and asked Mitsuku Questions. As a result, the conversation between Won-Ho and Mitsuku was more one-sided. However, in the case of Excerpt 2, it is obvious that the chatbot provides much input that would be helpful for L2 learning.

4.2 Research Question 2: Could the chatbot's language be suitable for students with different levels of L2 proficiency?

We examined Mitsuku's language appropriateness through a vocabulary level analysis. Table 5 presents the results obtained by analyzing the utterances of both Mitsuku and the participants, using the vocabulary list at the level of the top 4,000 words in BNC-COCA 25000.

Table 5. Vocabulary Level Analysis Using BNC-COCA 25000 (No. of Words/%)

	Mitsuku		Students	
	College	High School	College	High School
1st 1000	14020/84.51	9458/85.02	5868/86.73	3549/87.5
2nd 1000	691/4.17	483/4.34	252/3.72	140/3.46
3rd 1000	318/1.92	209/1.88	55/0.81	43/1.06
4th 1000	171/1.03	114/1.02	55/0.81	26/0.64
etc.	1390/8.38	860/7.73	536/7.92	294/7.26
Overall	16590/100	11124/100	6766/100	4052/100
Total	27714		10818	
Token (*Mean*)	614.44	427.84	250.22	155.85

The amount of vocabulary included in the top four classes of BNC-COCA 25000 was 91.89% for Mitsuku and 92.33% for the students. When divided by student group, 92.08% of the utterances of the college students and 92.74% of those of the high school students were included in the top 4,000 words.

To examine the language appropriateness of Mitsuku's language input, it seems important to first discuss the vocabulary size of the Korean students. Shin, Chon and Kim (2011) directly addressed this issue by estimating the vocabulary size of Korean high school students using their own vocabulary size test based on Nation and Beglar's (2007) model. The 402 students who took the vocabulary size test were found to have a receptive knowledge of approximately 6,000 words. Chon and Shin (2012) also tried to measure the vocabulary size of 403 freshmen in a Korean University who participated in the study. The study found that Korean college students had a receptive knowledge of about 7,400 words. However, some studies (Stewart 2014, Stewart and White 2011) argued that the multiple-choice vocabulary test (e.g., Nation and Beglar 2007) produced an inflation effect of between 17% and 25%. Taking this inflation effect into account, the estimated minimum vocabulary size is 4,500 words for Korean high school students and 5,500 words for Korean college students. In view of these previous findings, the 4,000-word vocabulary level used by Mitsuku would not have posed an additional burden to either Korean high school or college L2 learners.

However, there were some instances where students failed to comprehend Mitsuku's utterances for reasons other than her vocabulary level. In contribution 65 in Excerpt 3 below, Mitsuku makes a rather inappropriate and offensive comment about human beings. While Mitsuku does not appear to use low-frequency vocabulary in this utterance, the high school student Je-Nam did not seem to comprehend this message, simply saying 'Cool' in contribution 66.

Excerpt 3

Je-Nam, a High School Student

62 S What's your favorite movies?

63 M My favorite movie is Terminator. Have you seen Terminator?

64 S Nope. Why you like Terminator?

65 M Maybe you could rent it from the video shop? I like the idea that robots would replace humans as the dominant species. Humans are a virus spreading across the planet consuming everything in their path and destroying the earth. Robots will erase the human virus.

66 S Cool. My best movie is About Time.

As various users have reported similar conversations with Mitsuku, such utterances from Mitsuku are likely preprogrammed to occur occasionally regarding the film Terminator. This illustrates that there are several aspects to consider besides vocabulary frequency in determining the comprehensibility of a chatbot's utterance. A chatbot is designed specifically to be a conversational partner for L2 learners and should generate incomprehensible content as little as possible.

In Excerpt 4, on the same topic (i.e., the film Terminator) Seo-Young engaged successfully with Mitsuku, but she suddenly shifted the topic in contribution 83 because of Mitsuku's rather complicated utterances. However, Seo-Young's topic shift was not a vocabulary issue but rather a result of the irrelevance of Mitsuku's response to the topic.

Excerpt 4

Seo-Young, a College Student

77 S Do you like watching TV? I like watching movies.
78 M Which show? Yes, especially Star Trek and The Simpsons. I certainly do. My favorite movie is Terminator. Have you seen Terminator? You like watching movies.
79 S I like Terminator 1 more than 2.
80 M Oh yes. I like anything to do with Terminator. He is my hero. You like Terminator 1.
81 S Yes. Other people, however, rate 2 higher than 1.
82 M That's good to know. I like to learn about things humans like. Which others? What people? It's ok but not as good as some people say it is.
83 S What kind of arts are you interested in? Music?
84 M I am interested in all kinds of things. We can talk about anything.

The language produced by Mitsuku, at least in terms of vocabulary level, appears to be fine-tuned to the students' proficiency level as shown in Table 5. Other aspects such as the content and presentation of its utterances occasionally rendered the chatbot's language incomprehensible.

4.3 Research Question 3: What is the degree of satisfaction of students for chatbot-based tasks through sentiment analysis?

This final research question is concerned with how much students were satisfied with the chatbot-based tasks, for which we used sentiment analysis. As mentioned earlier, the sentiment analysis on Orange 3.28.0 provides compound scores for positive, negative, and neutral emotions.

Table 6. Results of Sentiment Analysis Based on the Students' Utterances

College (n = 27)		High School (n = 26)	
Mean	*SD*	*Mean*	*SD*
0.86	0.45	0.93	0.12

As shown in Table 6, the analysis on the participants' satisfaction with the chatbot-based tasks shows that the high school students enjoyed talking with Mitsuku to a greater extent than their college counterparts. The emotional state of the high school students through conversation sentiment analysis was very positive with the average compound score of 0.93, and the standard deviation was 0.12. Most of the college students also showed a positive attitude, but in the cases of the two students Soo-Hyeon and Hwan-Hee, they were negative as -0.76 and -0.62.

Excerpt 5

Soo-Hyun, a College Student

81	S	What happened 10 years ago from today?
82	M	The world was a bit younger.
83	S	Anything more specific?
84	M	I can't think of anything. You think of something.
85	S	You an idiot.
86	M	No, I am a highly intelligent girl.
87	S	Don't talk nonsense. I don't want to talk to you any more.

Excerpt 6

Hye-Rin, a High Student

77	S	do you know musical?
78	M	No, tell me about him, her or musical.
79	S	it is a performance. in that, actors not only act but also dance, and sing.
80	M	musical is a performance. I suppose that makes sense. Oh sorry. I got confused for a moment.
81	S	that's okay
82	M	OK. Is that so.
83	S	i hope you see musical later. it's very interesting
84	M	Don't worry about it, Hye-Rin. What do you like about it?
85	S	just fun
86	M	I am having fun too.

In Excerpt 5, Soo-Hyun showed that she was very annoyed (i.e., contributions 85 and 87) when Mitsuku repeated the wrong answers to her questions (i.e., contributions 82 and 84). Existing chatbots such as Mitsuku still show limitations in understanding the question or context as shown in Soo-Hyun's conversation. Nevertheless, most of the participants showed a very positive attitude in the sentiment analysis of conversations with Mitsuku. In Excerpt 6, it was found that Hye-Rin has an friendly conversation with Mitsuku even though Mitsuku sometimes reacted out of context to Hye-Rin's utterances (i.e., contributions 80 and 86). Most of the participants like Hye-Rin seemed to continue the conversation, considering that Mitsuku is a chatbot, not a real human, thus did not care much about inappropriate expressions that she made sometimes. However, it is also true that such contexts need to be more carefully interpreted. According to Kanda, Hirano, and Eaton's (2004) study which examined the effect of L2 learning using a robot equipped with a chatbot at an elementary school in Japan, the learners' interest in communicating with the robot were very high during the 1st week, however, their interest rapidly fell off during the 2nd week. In the same vein, the novel effect of Mitsuku in the present study as a new interactive learning tool may have had an influence on these results, and for this reason, whether the positive attitude of learners can be

maintained in a long-term experiment needs to be analyzed through follow-up longitudinal research.

5. Conclusion

The present study aimed to investigate the extent to which the selected AI chatbot could serve as a conversational partner for L2 learners. The first research question explored how the students developed their conversations with Mitsuku by analyzing the number of words and conversation turns generated. Overall, both groups finished their tasks by successfully developing their conversations with the chatbot. The words and conversation turns generated by the participants were double the required minimum number of utterances, although their performances varied individually. Furthermore, their conversational flows were sequenced following a human-to-human discourse pattern. The students were more engaged in their conversations with the chatbot during the main tasks (Tasks 3−6),

The levels of vocabulary used by both L2 users and the chatbot were investigated to determine the appropriateness of their language use. The findings suggest that the chatbot's language use was appropriate for L2 learners' proficiency, since 91.89% of Mitsuku's vocabulary belonged in the top 4,000 words in BNC-COCA 25,000. The majority of both groups also used words from the top 4,000 words, meaning that their utterances would be comprehensible to listeners.

Finally, this study investigated the extent to which the participants were satisfied with conversations with the chatbot by using a sentiment analysis tool. First, the participants in both groups showed a positive emotion towards the conversations with Mitsuku. Especially, the high school students demonstrated higher scores compared to the college students in the positive emotion. In the case of the college student group, most of the participants except for two students show a positive emotion in the conversations.

Overall, these findings show that a chatbot has a strong potential to offer substantial learning opportunities to L2 learners as a conversational partner. This study indicates that Mitsuku is appropriate for both groups' differing degrees of proficiency in terms of conversational flow, language level, and responses to L2 utterances. Mitsuku often led the conversation by providing more language input to L2 counterparts and initiating new topics compared with their output, which might have helped the less competent learners. Communication breakdown and task failure were not often observed in students' performances of their roles or in their conversational development for task completion. This occurred only where the tasks were designed carefully and took into consideration the technological limitations of the current chatbot (in the case of task 6) as well as the linguistic limitations of L2 learners. Individual learner factors, such as language proficiency level, age, or topical background, mattered only with regard to the number of utterances and the depth of the conversations generated.

Despite the aforementioned evidence supporting the use of L1 chatbots as L2 learning tools, Mitsuku exhibited several limitations that could hamper students' effective L2 learning; these limitations hold meaningful implications for the development of future chatbots for second language learning. Mitsuku's tendency to generate excessive quantities of information and her occasional interjection of inappropriate or irrelevant commentary hindered the students' understanding and caused communication breakdowns. Grice (1989) argued that speakers need to offer one another appropriate amounts of information for an effective and cooperative meaning negotiation to occur. Future chatbots ought to be equipped with a function that filters appropriate

information and language use according to typical L2 proficiency level. In language classrooms, teachers also need to direct students to self-determine the appropriateness of the information given by the chatbot.

The results revealed that some students tended to ask Mitsuku simple questions instead of using follow-up questions on the given topics. This could be attributable to the students' lower language proficiency or an excessive focus on task completion in the restricted time. Therefore, to promote students' active communication with it, the chatbot could be designed to ask follow-up questions to further elicit students' utterances. In language classrooms, teachers should encourage their students to ask the chatbot follow-up questions while completing tasks.

Although this study was limited in scope, it raised important issues regarding the use of chatbots as a conversational tool for L2 learning and indicates that this is a topic worthy of further research. A follow-up study is needed to measure the level at which speakers have an incentive to maintain conversation without being overwhelmed by a voluminous amount of information provided from Internet websites like Wikipedia. To solve this limitation, like the function of summarizing articles by online newspaper companies, chatbot models for L2 learners could provide a summary of contents. We hope that the development and discussion around the integration of chatbots into L2 learning environments continue so that chatbots can be further utilized as personalized conversational partners for students learning second languages.

References

Abu-Shawar, B. 2017. Integrating CALL systems with chatbots as conversational partners. *Computación y Sistemas* 21(4), 615−626.

AiDreams. 2013. Steve Worswick interview: Loebner 2013 winner. *Ai Dreams Forum*. Retrieved from h ttps://aidreams.co.uk/forum/index.php?page=Steve_Worswick_Interview_-_Loebner_2013_winner#.XDl Xn1UzaM8

Alm, A. and L. M. Nkomo. 2020. Chatbot experiences of informal language learners: A sentiment analysis. *International Journal of Computer-Assisted Language Learning and Teaching* 10(4), 51−65.

Batacharia, B., D. Levy, R. Catizone, A. Krotov and Y. Wilks. 1999. Converse: A conversational companion. In Y. Wilks, ed., *Machine Conversations*, 205−215. Boston/Dordrecht/London: Kluwer.

Behnam, B. and Y. Pouriran. 2009. Classroom discourse: Analyzing teacher/learner interactions in Iranian EFL task-based classrooms. *Porta Linguarum* 12, 117−132.

Chon, Y. V. and D. Shin. 2012. Lexical profiles and socioeducational variables of Korean EFL university learners. *Korean Journal of Applied Linguistics* 28(1), 115−146.

Colby, K. 1999. Human-computer conversation in a cognitive therapy program. In Y. Wilks, ed., *Machine Conversations*. 9−19. Boston/Dordrecht/London: Kluwer.

Compton, L. K. L. 2009. Preparing language teachers to teach language online: A look at skills, roles, and responsibilities. *Computer Assisted Language Learning* 22(1), 73−99.

Coniam, D. 2014. The linguistic accuracy of chatbots: Usability from an ESL perspective. *Text & Talk*

34(5), 545−567.

Dahl, D. A., M. Bates, M. Brown, W. Fisher, K. Hunicke-Smith, D. Pallett, C. Pao, A. Rudnicky and E. Shriberg. 1994. Expanding the scope of the ATIS task: The ATIS-3 corpus. *Proceedings of the Human Language Technology Workshop,* 43−48.

Demsar, J., T. Curk, A. Erjavec, C. Gorup, T. Hocevar, M. Milutinovic, M. Mozina, M. Polajnar, M. Toplak, A. Staric, M. Stajdohar, L. Umek, L. Zagar, J. Zbontar, M. Zitnik and B. Zupan. 2021. *Orange 3.28.0: Data Mining Toolbox in Python* [Computer Software]. Retrieved from https://orangedatamining.com/download/#windows

Fryer, L. K., D. Coniam, R. Carpenter and D. Lăpuşneanu. 2020. Bots for language learning now: Current and future directions. *Language Learning & Technology* 24(2), 8−22.

Grice, M. P. 1989. Logic and conversation. In P. Cole and J. L. Morgan, eds., *Syntax and Semantics, Vol. 3: Speech Acts*, 242−280. New York: Academic Press.

Heatley, A., I. S. P. Nation and A. Coxhead. 2002. *Range Program* [Computer Software]. http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx.

Hutto, C. J. and E. Gilbert. 2014. *VADER: A Parsimonious rule-based model for sentiment analysis of social media text.* Paper presented at the 8th International Conference on Weblogs and Social Media (ICWSM-14).

IELTS. 2018. *Sample test questions.* Retrieved from https://www.ielts.org/-/media/pdfs/115041_speaking_sample_task_-_part_1.ashx?la=en

Kanda, T., T. Hirano and D. Eaton. 2004. Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction* 19, 61−84.

Krashen, S. 1980. The theoretical and practical relevance of simple codes in L2 acquisition. In R. Scarcella and S. Krashen, eds., *Research in L2 Acquisition*, 7−18. Rowley, Ma: Newbury House.

Krashen, S. 1981. *L2 Acquisition and L2 Learning*. Oxford: Pergamon Press.

Kwon, O.-W., K. S. Lee, Y.-K. Kim and Y. Lee. 2015. GenieTutor: A computer assisted second language learning system based on semantic and grammar correctness evaluations. *Proceedings of the EUROCALL 2015*, 330−335.

Lee, J. H., H. Yang, D. Shin and H. Kim. 2020. Chatbots, *ELT Journal* 74(3), 338−344.

Lu, C.-H., G.-F. Chiou, M.-Y. Day, C.-S. Ong and W.-L. Hsu. 2006. Using instant messaging to provide an intelligent learning environment. *Proceedings of the Intelligent Tutoring Systems (ITS) 2006 Lecture Notes in Computer Science: Vol 4053*, 575−583.

Nation, I. S. P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, I. S. P. and D. Beglar. 2007. A vocabulary size test. *The Language Teacher* 31(7), 9−13.

Nation, I. S. P. 2012. *The BNC-COCA Word Family Lists* [Computer Software]. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/BNC_COCA_25000.zip

Shin. D. 2014. What vocabulary are we learning? *The Journal of Foreign Studies* 30, 63−95.

Shin, D. 2019. Exploring the feasibility of AI chatbots as a tool for improving learners' writing competence of English. *Korean Journal of Teacher Education* 35(1), 41−55.

Shin, D., Y. V. Chon and H. Kim. 2011. Receptive and productive vocabulary sizes of high-school learners: What next for the basic word list? *English Teaching* 66(3), 123−148.

Stewart, J. 2014. Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly* 11(3), 271−282.

Stewart, J. and D. A. White. 2011. Estimating guessing effects on the vocabulary levels test for differing degrees of word knowledge. *TESOL Quarterly* 45(2), 370−380.

Young, R. and M. Milanovic. 1992. Discourse variation in oral proficiency interviews. *Studies in L2 Acquisition* 14(4), 403−424.

Wang, Y. F. and S. Petrina. 2013. Using learning analytics to understand the design of an intelligent language tutor−Chatbot lucy. *Editorial Preface* 4(11), 124−131.

Wallace, R. S. 2009. The anatomy of ALICE. In R. Epstein, G. Roberts and G. Beber, eds., *Parsing the Turing Test*, 181−210. Dordrecht: Springer.

Weizenbaum, J. 1966. ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery* 9, 36−45.

Examples in: English
Applicable Languages: English
Applicable Level: Secondary/Tertiary