Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

2022, Volume 26, Issue 1
pp. 1–18

EMERGING TECHNOLOGIES

Design principles and architecture of a second language learning chatbot

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

Abstract

The purpose of this article is to set out the design principles and architecture of a second language (L2) learning voice chatbot. Building on L2 acquisition theories and chatbot research, in this article, we report on a South Korean government-funded longitudinal project in which we designed and developed a chatbot called "Ellie". Chatbot Ellie has three chat modes, "General Chat," "Task Chat," and "Skills". In the General Chat mode, L2 users can have short talks about personal information, whereas in the Task Chat mode, they can engage in a wide range of problem-solving L2 tasks to achieve task goals by exchanging meanings with Ellie. The Skills mode offers form-focused language practice. Ellie was piloted among 137 Korean high school students, who used Ellie individually or in a group, for seven weeks in their English classes. The quality of the chatbot was investigated in terms of the appropriateness of language level, continuity of conversation, and success in task performance. Based on the results of the pilot, Ellie appears to have considerable potential to become an effective language learning companion for L2 learners, and has implications for the design and developments of future L2 chatbots.

Keywords: *Artificial Intelligence, Chatbot, System Architecture, Task Design*

Language(s) Learned in This Study: *English*

APA Citation: Kim, H., Yang, H., Shin, D., & Lee, J. H. (2022). Design principles and architecture of a second language learning chatbot. Language Learning & Technology, 26(1), 1–18. http://hdl.handle.net/10125/73463

Introduction

A chatbot is a conversational user interface (conversational UI), which allows potential users to engage in a meaningful verbal exchange with a computer program (Lee et al., 2020; Wang & Petrina, 2013). These types of programs, built for human-computer conversational interaction, originated in the speech processing scientific community in the 1970s under the name of "dialogue systems" (Chen et al., 2017), and were further developed in the computer science community for text-based conversation. In the wake of rapid technological advances in the form of natural language processing (NLP) (Hirschberg & Manning, 2015; Zeroual & Lakhouaja, 2018) and deep learning (Dargan et al., 2020), chatbot communication is becoming increasingly popular for its capability to simulate human conversation (Chen et al., 2017; Dale, 2016). Although a human-like open conversation chatbot has yet to be developed, chatbots have been ubiquitously applied for the purpose of assisting humans in specific contexts, such as performing commands, answering questions, or providing information (Diederich et al., 2019). Moreover, the advent of user-friendly chatbot building platforms, such as Google Dialogflow™ and IBM Watson™ has accelerated the growth and expansion of chatbot services to a wide range of fields.

In the field of computer-assisted language learning (CALL), chatbots have recently started to attract interest among the relevant research and pedagogical communities as potential language learning resources (Bibauw et al., 2019; Fryer et al., 2020), with considerable possibilities for use in outside-classroom language practice (Alm & Nkomo, 2020). Most studies concerning CALL have been dedicated to examining the potential of first language (L1) chatbots (i.e., those originally built for native speakers of that language), with only a few attempts having been made to design and develop a chatbot exclusively intended for second language (L2) learning. One such attempt, namely, the Dave E.S.L chatbot (see Coniam, 2004 for a review of this chatbot), was developed from a pilot project using Artificial Intelligence Markup Language (AIML), with technical support provided by the Artificial Linguistic Internet Computer Entity (A.L.I.C.E.) Foundation. However, this ESL tutor bot has received little attention from the L2 research community in terms of evaluating its effectiveness for L2 learning. Despite several other related efforts within the L2 research and pedagogical community, including the development of the Lucy chatbot (Wang & Petrina, 2013), the Dialog-Based CALL system (Huang et al., 2017), and Mondly (see Fryer et al., 2020 for an introduction to this chatbot), there has been little progress in the development of L2 conversational UI, especially in terms of using interlanguage data related to L2 learners' language behaviors and considering second language acquisition (SLA) frameworks.

The development of an effective L2 chatbot has been rather slow primarily because developing an L2 chatbot involves considerably greater challenges than developing its L1 counterpart. An L2 chatbot should not only ensure that a human-computer conversational interaction is possible, but also function as a language tutor to assist L2 learners in communicating, and provide appropriate activities or references for successful L2 learning. To develop such a chatbot requires knowledge concerning appropriate design and architecture systems, and in defining (chatbot) agent types, as well as an informed understanding of SLA theories and instructional approaches. Furthermore, developers of an L2 chatbot need to have an adequate knowledge of the target L2 learners' interlanguage, language learning needs, and proficiency (Fryer et al., 2019).

In view of the need for the development of a chatbot exclusively designed for the purpose of L2 learning and teaching, in this article, we introduce a novel approach for creating an artificial intelligence (AI) chatbot geared toward enhancing oral communication skills among L2 learners, for reference in future research. In a two-year longitudinal South Korean government-funded project, an L2 learning task-based chatbot known as "Ellie" was designed in accordance with SLA theories, and developed and evaluated to deal with the current technological limitations of existing chatbots. Based on our experience gained in the construction of Ellie, in this article, we aimed to set out a system architecture and task types in relation to an L2 learning chatbot in accordance with recent CALL developments, and to assess the quality of this chatbot by analyzing its language and the task performances of a select group of L2 users.

Conversation Systems in AI Chatbots

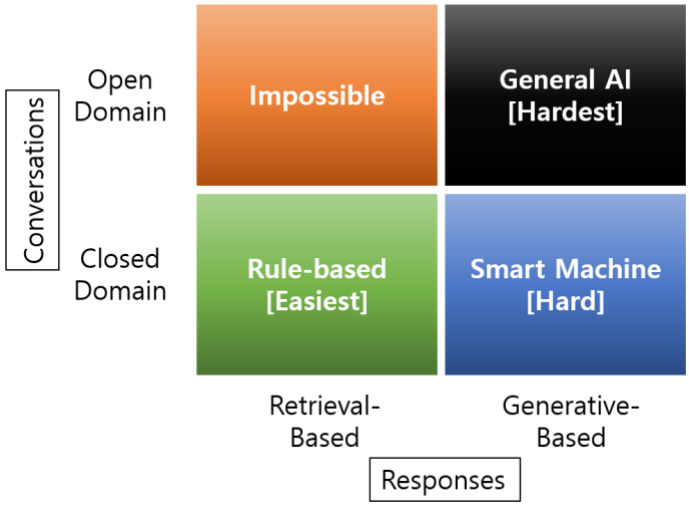
In this section, we discuss the taxonomy of chatbot models, the system architecture for chatbots, and data generation and management related to chatbot language. Consideration is also given to designing chatbot-based tasks for L2 learning and practices.

Taxonomy of Chatbot Models

Chatbot users expect responses from the relevant interlocutor to be free-flowing and accurate, without being made aware of computational complexity or being presented with ambiguity, and the underlying program needs to satisfy these conditions. Figure 1 illustrates Kojouharov's (2016) taxonomy of a chatbot model, which elucidates the challenges in achieving free-flowing and accurate responses.

Figure 1

Taxonomy of a Chatbot Model (adapted from Kojouharov, 2016)



As can be seen in Figure 1, chatbot conversation can occur in an open or a closed domain. In the open domain, the conversation has no direction, similar to human-to-human conversational interaction; hence, users are expected to engage in extended conversations on a wide range of topics. Building such a chatbot would require a large amount of language data and sophisticated learning algorithms related to diverse topics and world knowledge, which makes it a daunting task. Meeting the requirements of the Turing test (Turing, 1950) and the Loebner prize (Bradeško & Mladenić, 2012) have stimulated institutional efforts toward achieving human-likeness using such open-domain chatbots for several decades. In contrast, in the closed domain, a conversation is more goal-oriented and functional, with domain knowledge, topics, and contexts for conversation being predefined. In such a domain, the conversational UI is intended to fulfill the programmed tasks as efficiently as possible. Thus, the resulting dialogue could be sustained only in terms of the predesigned tasks and topics, but the recognition rate is expected to be higher than that in open-domain chatbots under the same conditions. Assistant bots and call center bots are examples of chatbots in this domain (Shum et al., 2018).

Another dimension within this taxonomy relates to the nature of chatbot responses, which can be distinguished as retrieval-based or generative-based chatbot responses. A retrieval-based chatbot responds to users through already-scripted rules. The responses can be generated either through rule-based intent matching or machine learning training. Conversely, a generative-based chatbot produces new responses from scratch. Generative models are typically based on deep learning and machine translation techniques, which are difficult to achieve. Hence, open-domain generative-based conversational chatbots (i.e., general AI in Kojouharov's taxonomy) have not yet been successfully developed, although the recent AI developments of big language models such as GPT-3 (Brown et al., 2020) and Switch-C (Fedus et al., 2021) may advance the realization of general AI, with its promise of open domain functionality. Nevertheless, the language output of these models (or AI-based systems; e.g., GPT-3 and Switch-C), if they were to be applied to the developments of chatbots for language learners, should be carefully monitored, as the text-based datasets (mostly from the Internet) on which they are trained may "overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations" (Bender et al., 2021, p. 610; see Godwin-Jones, 2021 for a discussion on a similar issue in the context of L2 education).

According to a survey conducted by Abdul-Kader and Woods (2015), Loebner prize winners such as A.L.I.C.E versions (years 2000, 2001, 2004) and Mitsuku (years 2013, 2016, 2017, 2018, 2019) were built based on advanced pattern matching in AIML. The AI chatbots such as Google Assistant and Alexa,

respectively installed in Google Home and Amazon Echo, can be categorized as closed-domain retrieval-based chatbots. Most of their conversations end in limited one-turn dialogue, but their responses to information requests are almost unlimited, due to web scraping, natural language understanding, and other algorithms. Finally, Microsoft's recent social chatbot, XiaoIce (Zhou et al., 2020), was introduced as an open-domain conversation chatbot for Chinese speakers, which seems closest to "General AI."

In this project, we aimed to develop a closed-domain retrieval-based chatbot, as our target chatbot was intended to facilitate the engagement of L2 learners in predefined tasks on a limited range of topics (i.e., closed-domain), and its responses were carefully written by the developers (i.e., retrieval-based) in terms of specific L2 learners' proficiency levels so that these target L2 learners would be exposed to more comprehensible L2 input from chatbot-learner conversational interaction.

System Architecture

Before developing a target application, it is important to design its architecture, which underpins "the understanding, implementation, maintenance, and further development" of a chatbot system (Khan, 2017, p. 114). According to Khan (2017), a scalable, sustainable, and standardized architecture is essential for successful chatbot development. He further elucidated on standardized chatbot architecture and user interface components, and subsequently demonstrated the differences involved in the implementation of differing bot types. A typical chatbot solution architecture comprises multiple layers: presentation, business, service, data, and utility (Khan, 2017). First, the presentation layer displays the user interface and controls user interaction. In the business layer, data are processed and formatted, and dialogues are encoded into a set of data structures. The service layer components provide access to both internal and external data and NLP services, while the data layer concerns data collection and storage that need to be carefully designed for future analysis. Finally, the utility layer concerns security, logging, and configuration.

The fundamentals of system architecture are adjusted and implemented according to the role and types of bot design. Khan (2017) set out how generic standardized architecture could be applied to personal assistant, customer service, and functional bots. In line with Khan's thinking, some recent research (e.g., Di Prospero et al., 2017) has started to offer system architecture in relation to newly developed chatbot agents. In the present study, we also referred to Khan's multi-layered model as well as to example architectures of personal assistant and customer service bots for the development of a more standardized and sustainable chatbot design.

Data Generation, Analysis, Training, and Accumulation

Within the architecture system of conversational UI, a user's speech input (i.e., voice data) is recognized through automatic speech recognition (ASR) technology and transformed into text data. Rule-based (i.e., data-driven) chatbots then understand the text by matching it to prefabricated conversation pair datasets (i.e., intents), and retrieve and respond through their respective dialogue managers. With chatbot responses being stored in text format, they can be synthesized with corresponding audio files, which creates speech output. In this system of rule-based chatbots, a very large amount of data needs to be generated, analyzed, trained, and accumulated by developers for these chatbots to maintain human-like conversations. This is in contrast with generative-based chatbots, which do not rely on an already established dataset, but are programmed to automatically generate output responses word by word via deep learning, such as the encoder-decoder deep neural network model (Shum et al., 2018) and long short-term memory (Hochreiter & Schmidhuber, 1997), which facilitate visual awareness, sentiment detection, topic guidance, or personality consistence (Shum et al., 2018).

Fundamentally, there are two types of data involved in conversational UI within this system: structured and unstructured. Structured data includes numerical or fixed data, such as names, dates, addresses, or some types of established short conversations. Structured data can be easily organized and understood by the computer program. In contrast, unstructured data are mostly conversational data made up of text, video, or audio, which are difficult to deconstruct and organize. Since most chatbot data are unstructured, they need to be predefined and trained (Greyling, 2019). Consequently, data training is essential for building an

effective chatbot. A typical training process in this regard involves machine learning, whose effectiveness is predicated by the amount of input data fed for the training: the more trained data it accumulates, the smarter the chatbot becomes.

The latest chatbot builder application programming interfaces (APIs) such as Dialogflow™, which are equipped with machine learning algorithms, entity detection, and NLP, help to organize chatbot data more efficiently and automatically. Since chatbot builders effectively assist in the entire process of chatbot building, developers can significantly reduce the time and technical effort needed for smooth human-computer interaction; specifically, they can channel their energy into determining what intents or entities are needed, which decision-making task structures are appropriate for the target conversational agent, and how data should be fed into the system architecture. In view of such advantages, we used Dialogflow™ for the purpose of data generation, analysis, training, and accumulation (see the System Architecture of Ellie section for details).

Chatbot and Second Language Tasks

Chatbots have been suggested as useful for L2 learning and practice, with L2 learners' preference for chatbots over human partners having "its source in the fear of making mistakes and appearing less than competent" when interacting with human partners (Fryer et al., 2020, p. 16). A recent study conducted by Fryer et al. (2019) revealed that L2 learners also appreciated the chatbot's capability to expose them to a wide range of conversational expressions and vocabulary, and "enable repetitive practice" (p. 286). However, it is unlikely that a chatbot would function as an effective L2 learning partner in the absence of appropriate L2 tasks and associated guidance by instructors. Considering this point, Kim (2018) proposed a need for chatbot-based L2 tasks.

A major rationale for requiring chatbot-based L2 tasks lies in the characteristics of existing tasks or activities implemented in already-available conversational UI (e.g., Alexa, Google Assistant), which are generally not suitable for L2 learning requirements, although some of them may appeal to L2 instructors and learners (see Lee et al., 2020 for some examples of how to use already-available conversational UI for L2 lessons). It has been contended that CALL tasks should be designed primarily based on the identification of L2 learners' needs (Chapelle, 1998; González-Lloret, 2003).

In designing chatbot-based L2 tasks, we referred to Chapelle's criteria for CALL task appropriateness (i.e., language learning potential, learner fit, meaning focus, authenticity), which we discuss below. First, language learning potential should be considered in designing CALL tasks, with some of the relevant components including interactional modification, modified output, time pressure, modality, and support, in line with task-based language teaching (TBLT) literature (Ellis, 2003; Nunan, 2004; Skehan, 1998; Willis & Willis, 2001). Second, the tasks should be designed in view of the learner fit, taking learner characteristics, such as language proficiency, age, learning style, and other important individual differences into consideration. Third, the tasks should be organized to focus primarily on accomplishing social goal-oriented tasks, such as decision making, information exchange, and problem solving. Finally, CALL tasks should have some level of authenticity, indicating that they should be similar to real-life tasks that learners are likely to encounter outside the classroom. These criteria should be applied when designing chatbot-based L2 tasks, so that L2 learners can acquire language skills effectively and sustain their interest in interacting with the chatbot.

Presenting Ellie, a Task-Based L2 Chatbot

This section presents the design principles and system architecture of the developed chatbot (i.e., Ellie). An introduction to Ellie's different chat modes is provided subsequently.

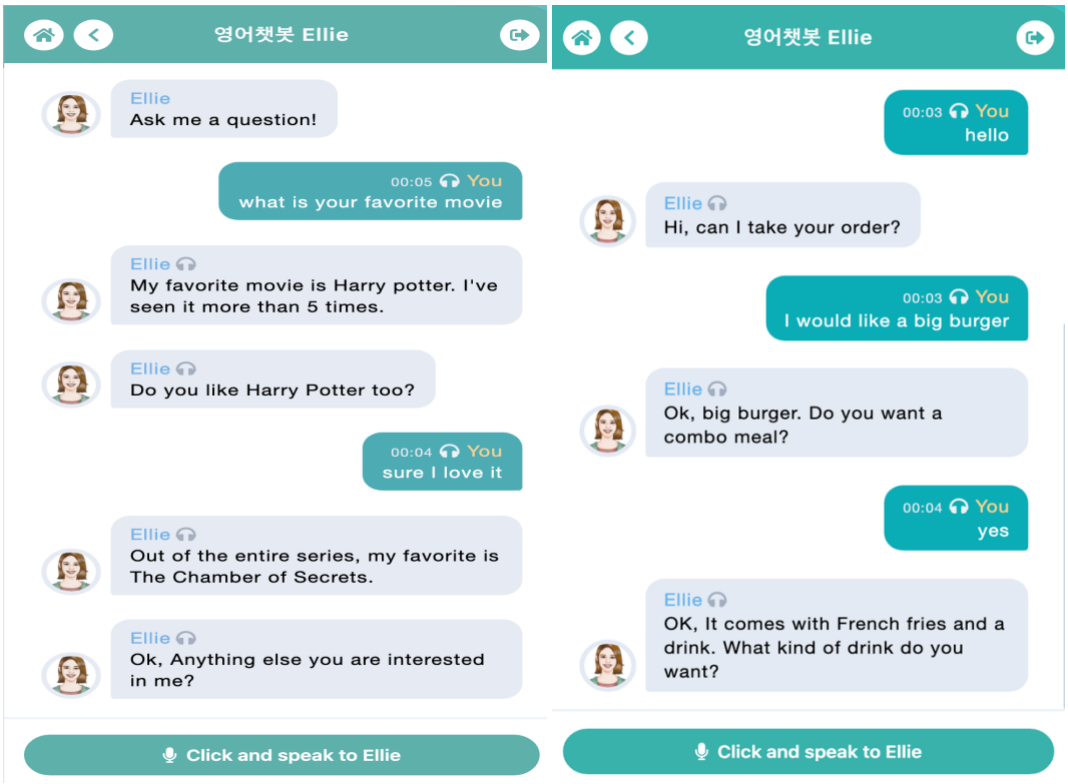
Design Principles

The first principle considered in the development of Ellie was that an L2 chatbot needs to provide both open and closed chat opportunities for learners. Open chat differs from open-domain chat, which was

discussed previously in reference to Figure 1. Open chat ("General Chat", henceforth) is more about exchanging personal information, through which L2 learners freely initiate conversation on predefined topics, and become acquainted with Ellie, in a natural conversational style. However, unlike open-domain chat, the conversation processes are often planned or structured, so the chatbot may not be able to respond to the user when conversation develops in wider topics. In contrast, with the closed chat ("Task Chat", henceforth), users can interact with the chatbot to engage in L2 problem-solving tasks (see the Three Chat Modes in Ellie section for a more detailed description). Figure 2 presents examples of General Chat (Left) and Task Chat (Right) between users and Ellie. In General Chat, the user asked about Ellie's hobbies and answered Ellie's questions to maintain the conversation. In Task Chat, a student was asked to select a type of hamburger, to respond to a clerk's (i.e., Ellie's) question, and to select a drink and payment method during the conversation.

Figure 2

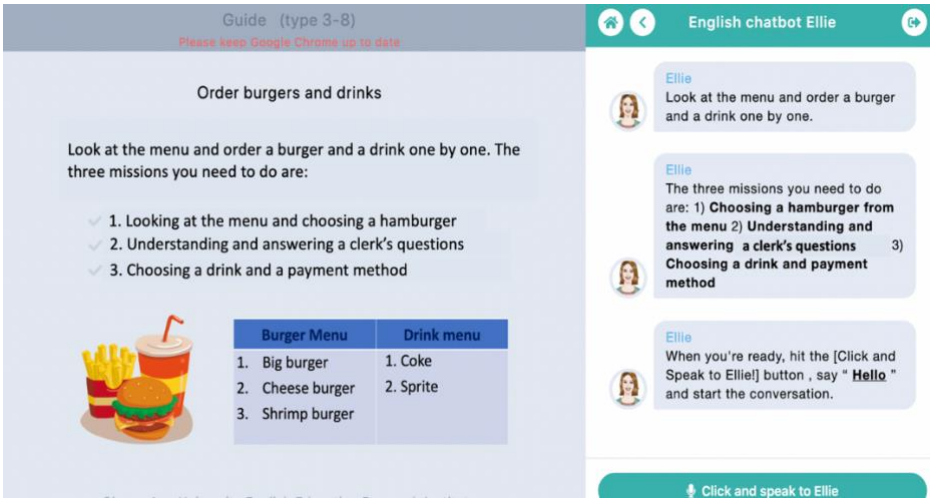
Examples of General Chat (Left) and Task Chat (Right)



The second principle concerned the need for L2 chatbot tasks to foster goal-oriented communication, as suggested in the TBLT literature (e.g., Nunan, 2004; Skehan, 1998; Willis & Willis, 2001). With this point in mind, we designed each task so that the initial page provides a clear direction concerning a task goal (i.e., the mission) and the procedure to follow to achieve that goal. In this way, users can initiate communication after being fully prepared beforehand to converse in English with Ellie. Visual clues (i.e., images and texts) have also been included for some tasks to reduce communication breakdowns that may arise due to non-linguistic elements. Figure 3 illustrates the visual clues and written instructions presented for one of the problem-solving tasks (ordering at a fast-food restaurant).

Figure 3

Example of Visual Clues Relevant to a Task



The third principle concerns the need for some degree of freedom in the chatbot-based tasks, which involves a divergence from traditional scripted dialogue or situational drills (Paulston, 1970). It was important for the tasks to be designed such that they were structured to some extent to ensure that conversations with Ellie were not fundamentally open-domain chat, with some options provided for learners to take or choose from when engaging in each task, which was intended to add interest and authenticity to the task. In doing so, learners may need to go through meaning negotiation processes involved in a task-based chat. For example, Excerpt 1 presents an example of a conversation between a high school student and Ellie. This excerpt demonstrates how a student accomplished a goal, which was to buy new shoes, through conversing with Ellie. In lines 5 and 9, the student selected a certain color for the shoes, but they were not available at the store, as Ellie responded in lines 6 and 10. Thus, the student kept suggesting different colors to find an available pair in that color at the store, which replicated a possible real-life situation. For some tasks, learners are also required to draw on their cognitive abilities, such as problem solving, strategic planning, or decision making to successfully achieve goals, which makes chatbot-based tasks qualitatively different from situational drills.

Excerpt 1.

1	S:	I am looking for new shoes
2	E:	Sounds good. What is your size?
3	S:	290
4	E:	Great. What color do you want?
5	S:	Black
6	E:	Oh, sorry. They are out of stock. What about other colors?
7	S:	Beach (The student mispronounced the word 'Peach').
8	E:	One more time?
9	S:	Yellow
10	E:	Oh, sorry. The last in that color are already reserved. What about other colors?
⋯ (student kept trying a different color) …		
15	S:	Brown
16	E:	Sure, here you are.
17	S:	How much are they?
18	E:	They're $95. Would you like to buy them?

Note. S: Student, E: Ellie

The final principle concerns the establishment of a chatbot persona. Given that recent research (e.g., Greyling, 2019; Ó Broin, 2017) has indicated the need for defining a distinct and recognizable persona with a human name for the chatbot interface, our conversation UI has been anthropomorphized; the chatbot was named, visualized with an illustrated character, and depicted as a female ESL teacher who lives and works in Korea. The Ellie profile further includes details concerning her family and educational background, living experiences, personality, and a certain look (see Figure 4 for Ellie's look). It was expected that using such a consistent persona in the conversational UI would give L2 users a feeling of interacting with a real person, with Ellie's utterances also being written as culture-coded (as someone growing up in San Francisco in the U.S.) and context-specific.

Figure 4

Ellie's Look

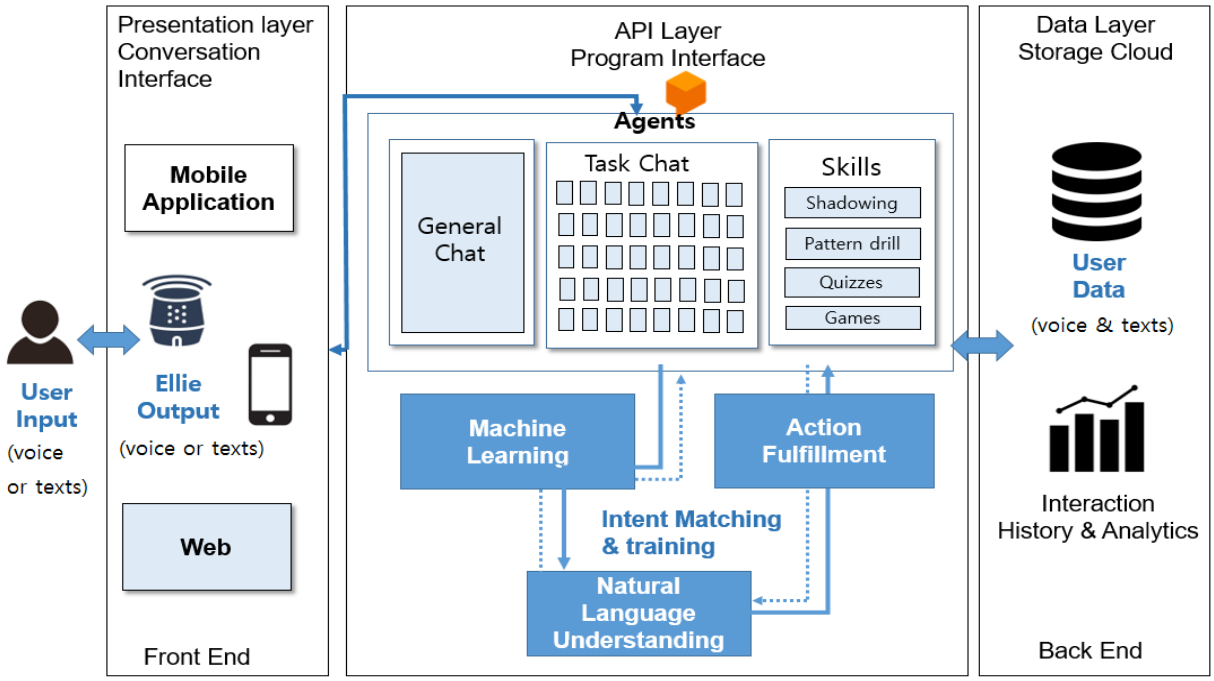
System Architecture of Ellie

Ellie was developed using a natural language API platform, Dialogflow™, which is Google's natural language understanding tool for building AI chatbots. On the Dialogflow™ platform, the chatbot agent can be created and trained, with steady improvements to competence by matching a user's data to already generated intents and entering training phrases (i.e., expected user utterances). By integrating technologies in the cloud platform, such as ASR and text-to-speech, machine learning algorithms allow rapid processing to comprehend the user's natural language.

Figure 5 illustrates the system architecture of Ellie, which comprises a presentation layer, an API layer, and a data layer. The data are generated in two tracks. First, chatbot dialogue and expected user utterances are manually coded in the program interface (i.e., in the API layer) through 65 agents (i.e., 1 for General Chat, 60 for Task Chat, 4 for Skills at present) and 1945 intents (see the Three Chat Modes in Ellie section for details). It was expected that predefining user utterances would enable the agent to more easily recognize and match user utterances to the predefined intents (and provide appropriate responses). The other data source is user input. The user's utterances are automatically converted into transcript through Google's Chrome Web API Speech Recognition at the conversation interface (i.e., in the presentation layer), and saved in data storage (i.e., in the data layer). As the transcript of the utterances are synchronously presented on the screen, the users can attempt to speak again if they notice an error in their pronunciation. Once the user input is entered, the agent detects and matches the data to the closest intent and transmits a response to the user. User logs are later analyzed and used for chatbot training through machine learning algorithms.

Figure 5

System Architecture of Ellie



Three Chat Modes in Ellie

In this section, the three chat modes in Ellie (i.e., General Chat, Task Chat, Skills) are discussed. Each mode was built for different L2 learning purposes. The first mode, General Chat, provides opportunities for L2 learners to ask Ellie about her personal information, interests, and thoughts, through which users are expected to perceive their conversation flow as normal informal communication. Currently, the agent of the General Chat has 1195 intents to maintain the context of the conversation. The topics of the intents in this mode were identified considering "learner fit" and "authenticity", which are essential criteria in terms of CALL task appropriateness (Chapelle, 2001), as noted previously. In this mode, L2 learners can practice daily conversation expressions they have learned previously by exchanging meanings with Ellie, which is a qualitatively different exercise from practicing these expressions through pre-scripted dialogue or role plays taken from textbooks. Ellie was initially designed to maintain only one or two turns per sub-topic, as is done with other assistant bots (e.g., Alexa). However, we have kept extending topic trees through follow-up phrasing so that users could maintain sustained conversations with Ellie, which is an important feature to consider in developing chatbots for language learning (Fryer et al., 2020).

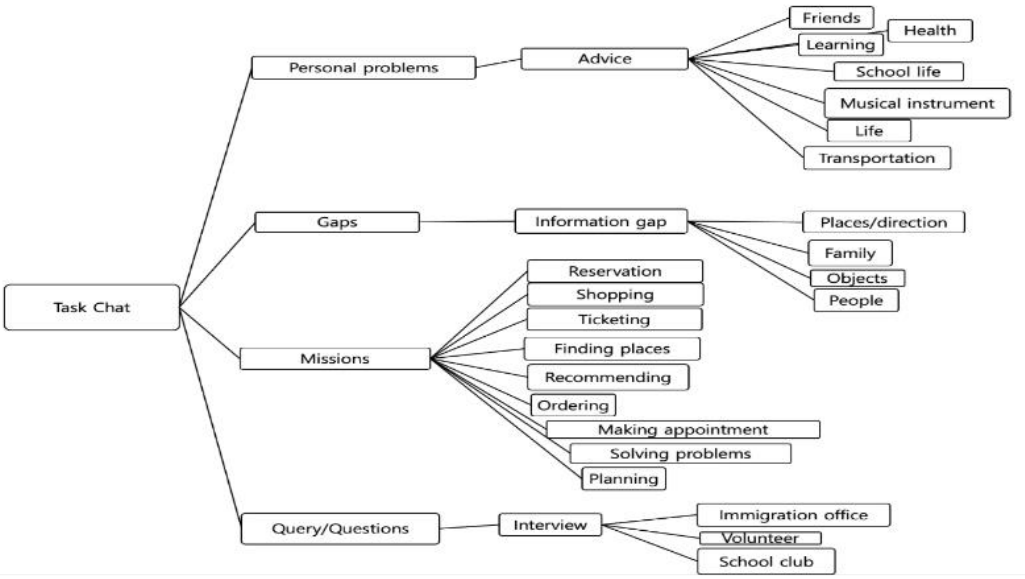
The second chat mode, Task Chat, which comprises a major portion of Ellie, is designed for interaction with a chatbot to complete L2 problem-solving tasks that involve goal-oriented and meaning-focused cognition activities, as recommended in the TBLT literature (Ellis, 2003; Nunan, 2004; Skehan, 1998; Willis & Willis, 2001). In this mode, users can practice a wide range of problem-solving tasks, which are further classified into four sub-types: (a) Personal Problems, (b) Gaps, (c) Missions, and (d) Query or Questions. Figure 6 illustrates the semantic map of these problem-solving tasks. Task Chat currently consists of 60 agents, each of which corresponds to one problem-solving task to be completed in 5–12 conversation-turns, depending on the topic and complexity of the tasks.

Each task and its (expected) conversation pattern between the chatbot and users were designed and visualized using decision trees (DTs) (Shah et al., 2018), which could be considered as providing a systematic master plan scenario of users' task completion. Lobo (2017) suggested that using DTs is the

most effective way to identify the intents for predicting users' utterances and to provide the appropriate responses. Using DTs, we aimed to predefine the conversation tree starting from the root question, which would be followed along hierarchical branches that narrow down to the user's goal through chatbot intents. With each task, the users are expected to become aware that they can freely exchange their own meanings with the chatbot, although they are directed toward achieving a predefined goal. Each task has similar DTs, but with differences in terms of task complexity. User direction and image clues are presented before a task begins (see Figure 3 for an example), which is an important strategy to define and restrict the scope of the conversation.

Figure 6

Organization of Task Types



The third chat mode, Skills, includes form-focused language practice through interactions with the chatbot. While "skills" as a term is generally used for apps in conversation interfaces, such as XiaoIce and Echo (Zhou et al., 2020), the term has been used to denote form-focused L2 practice in this study. Conversation with the chatbot in this mode is not geared toward meaning-based conversation, but rather for users to practice their pronunciation and vocabulary or play game-like language activities for speech enhancement. Skills has four agents: (1) shadowing, (2) pattern drills, (3) quizzes, and (4) games. Shadowing consists of repeat-after-me type exercises. If a user copies Ellie's target expression (i.e., a word or a sentence) accurately, the chatbot provides positive feedback on the accuracy of speech, and presents the next target expression. If a user fails to copy the target sentence accurately, the chatbot repeats that expression. In pattern drills, the user is asked to copy model situational dialogues (3–4 turns) consecutively in the same manner. Quizzes and games consist of the "listen, think, and answer" mode that includes memorizing, ranking, categorizing, matching, and guessing games.

Quality Assessment

In this project, piloting was conducted twice for quality assessment of Ellie. The quality of Ellie was assessed based on three aspects of L2 chatbot development. First, the appropriateness of language levels was evaluated by analyzing chat logs exchanged between Ellie and the users. Second, the continuity of conversation dialogue was investigated by counting conversation-turns per session (CPS). Lastly, the estimating task success rates (TSRs) was used to evaluate how successfully the expected meaning exchanges were fulfilled in the task-based chat. In this article, we report on the results of the second and more extensive pilot of Ellie on all the problem-solving tasks, as shown in Figure 6. The results of the first

pilot with high school and elementary school students on the three initially developed chatbot-based tasks have been reported elsewhere (see Yang et al., in press).

Participants and Test Procedure

A total of 137 high school students in South Korea have used Ellie. They talked with Ellie individually or in groups under a teacher's guidance for seven weeks. Due to the class schedules, they talked with Ellie for three weeks in the first semester and four weeks in the second semester. Each class lasted 50 minutes and the students talked with Ellie for 20 to 30 minutes per class using their own mobile phones. Depending on the class schedule of each day, the teacher sometimes assigned particular tasks to the students to practice or let them choose tasks to talk through with Ellie.

In the pilot, the students exchanged a total of 2,365 conversation sessions (i.e., the period from the start to the end of a user's talk with a chatbot, Shum et al., 2018) with Ellie, with a varying number of conversation sessions observed for each chat mode. The Task Chat mode (1917 sessions) was used most frequently, followed by the General Chat mode (420 sessions) and the Skills mode (28 sessions).

Appropriateness of Language Levels

Existing chatbots, such as A.L.I.C.E., Mitsuku, and Cleverbot were built for communicating with native English speakers, so the vocabulary level of those chatbots has not been controlled for the purpose of L2 learning. In phrasing Ellie's utterances, the vocabulary level of its language was taken into consideration, with excessively difficult vocabulary being removed, so that our target learners could be exposed to more readily comprehensible English input.

Table 1 shows the lexical profiles of Ellie's language based on the first 4,000 of the 25,000 words extracted from the British National Corpus and the Corpus of Contemporary American English by Nation (2016). The reason for applying the first 4,000 words to the analysis of the chatbot language is that the vocabulary level required for the English test for college admission in South Korea does not exceed this level of 4,000 words (Joo, 2008). The analysis was conducted for the General Chat and Task Chat modes. In addition, we present our results related to the Task Chat mode in terms of three different levels, as we designed the tasks such that the coverage of the first 1,000 words in the lower level tasks would be higher.

Table 1

Lexical Profiles of Ellie's Language in the General Chat and Task Chat Modes

Word Band	***General Chat***	***Task Chat***		
		Level 1	**Level 2**	**Level 3**
	Token/**%**	Token/**%**	Token/**%**	Token/**%**
1st 1000	19085/88.58	2922/91.08	3650/90.10	803/86.72
2nd 1000	847/3.93	119/3.71	160/3.95	76/8.21
3rd 1000	225/1.04	9/0.28	21/0.52	9/0.97
4th 1000	168/0.78	12/0.37	30/0.74	8/0.86
Proper Noun	520/2.41	78/2.43	15/0.37	6/0.65
No./Alphabet	53/0.25	24/0.75	35/0.86	3/0.32
Sub-total	20898/**96.99**	3164/**98.62**	3911/**96.54**	905/**97.73**
Others	647/3.00	44/1.37	140/3.46	21/2.27
Total	21545	3208	4051	926

Note. Level 1 requires 3 goals to be completed and 5–6 conversational turns in length, Level 2 requires 3 goals and 6–8 turns, and Level 3 requires 4 goals and 6–8 turns.

Although the Task Chat mode consists of three differing task levels, there was little difference in the coverage of the first 4,000 words in terms of the chatbot language. In addition, the coverage of those 4,000 words in the General Chat mode was also approximately 97%, which is almost identical to that of the Task Chat mode. The coverage of the first 1,000 words in Level 3 of the Task Chat mode was 86.72%, which was relatively lower compared to the other two levels. However, when compared to findings in a study by Nation (2006), it was still higher than that of the first 1,000 words in normal speech, with 81–84% of coverage. It was further found that the rate of use of low frequency vocabulary increases marginally as the task level increases. That is, the rate of the 4th 1000 band, which consists of relatively low frequency words compared to the other three word bands, was 0.37% in Task Level 1, but increased to 0.74% in Level 2 and 0.86% in Level 3.

Next, the lexical profiles of four different types of Task Chat modes (i.e., Personal Problems, Gaps, Missions, and Query or Questions) were analyzed, as shown in Table 2.

Table 2

Lexical Profiles of Ellie's Language in Different Task Types in Task Chat

Word Band	***Personal Problems***	***Gaps***	***Missions***	***Query/ Questions***
	Token/**%**	Token/**%**	Token/**%**	Token/**%**
1st 1000	858/93.87	1858/90.77	3951/89.92	1641/78.18
2nd 1000	29/3.17	59/2.88	203/4.62	176/8.38
3rd 1000	5/0.55	3/0.15	28/0.64	40/1.9
4th 1000	2/0.22	5/0.24	28/0.64	33/1.57
Proper Noun	0/0.00	60/2.93	34/0.77	45/2.14
No./Alphabet	7/0.77	1/0.05	46/1.05	28/1.33
Sub-total	901/**98.58**	1986/**97.02**	4290/**97.64**	1963/**93.5**
Others	13/1.42	61/2.98	104/2.37	136/6.48
Total	914	2047	4394	2099

The results showed that approximately 90% of the chatbot language involving the three task types, "Personal Problem," "Gaps," and "Missions," were from the first 1,000 words. The text coverage of the four levels of vocabulary was also 97% or more, which points to its suitability for EFL learners.

Continuity of Conversation

The conversations between Ellie and users were reviewed using CPS, as shown in Table 3. CPS indicates the average number of conversation exchanges between a user and the chatbot in a conversation session. A higher CPS means a user has actively engaged in conversation with the chatbot, showing the chatbot's capability in maintaining conversation with the user (Shum et al., 2018; Zhou et al., 2020). The results of this analysis revealed that conversation lasted up to 9.3 turns on average, indicating that the high school students maintained their conversation for much longer than other classroom conversation tasks (e.g., scripted dialogues or semi-structured speaking tasks) of normally 3-5 turns in the English curriculum or in textbooks. In terms of each chat mode, the students exchanged the highest CPS in the General Chat mode (mean: 10.7 CPS), followed by the Task Chat mode (mean: 9.6 CPS), and the Skills mode (mean: 5.5 CPS). For each task type in the Task Chat mode, the students showed the highest CPS when completing Missions (mean: 10.5 CPS), followed by Query/Questions (mean: 9.8 CPS), Gaps (mean: 8.8 CPS) and Personal Problems (mean: 7.3 CPS).

Table 3

Summary of Conversation-Turns Per Session (CPS) by Chat Mode

Chat modes	Task types	Mean	SD	Min	Max
General Chat	-	10.7	10.69	1	94
Task Chat	Personal Problems	7.3	2.72	2	29
	Gaps	8.8	5.39	1	94*
	Missions	10.5	2.80	2	47
	Query/Questions	9.8	5.57	2	31
	Sub-total	9.6	3.63	1	94
Skills	-	5.5	4.37	1	23
Total Average		9.3	3.76	1	94

Note. *For one of the Gap tasks in the Task Chat mode, one student continued to ask questions and self-correct their talk to complete the given task, subsequently producing 94 turns.

Task Success

The TSRs of chatbot users can also determine the quality of the chatbot (Shum et al., 2018), particularly in terms of successful meaning exchanges to complete tasks. The TSRs were calculated for Personal Problems, Missions, and Query/Questions. For these tasks, the students' task success was automatically checked and recorded in the database. However, another task type (i.e., Gaps) was excluded in this analysis because the students' task success for this task type was largely determined by the students' own decisions, and was not recorded in the chatbot system.

In analyzing the data to assess TSRs, certain invalid conversation sessions were carefully reviewed and removed from the data because communication failures in these sessions resulted from technical or environmental factors, such as failure of speech recognition due to noise or abrupt termination for unidentified reasons, rather than user factors or task design problems. In the three task types in Task Chat mode, 1,012 invalid conversation sessions, comprising 55.1% of the total conversation sessions in this mode (1,836 sessions) recorded in the user logs, were excluded.

Successful conversation sessions consisted of the sum of sessions where the students fully completed each task type.

$$task\ success\ rates = \frac{\text{successful conversation sessions}}{\text{total conversation sessions} - \text{invalid conversation sessions}}$$

Table 4 presents the students' TSRs by each task type and by each level in the Task Chat mode. Overall, moderately high success rates (65.5%) were observed on average for the three task types in Task Chat. In terms of each task type in the Task Chat mode, the students presented the highest TSRs for Missions (67.4%), followed by Personal Problems (60.7%), and Query/Questions (43.8%). The TSRs were also calculated in terms of each level. The sub-tasks in the Task Chat mode were categorized by difficulty level, ranging from Level 1 as the easiest to Level 3 as the most difficult. The results show that the highest TSRs were found in Level 1 tasks (69.9%), followed by Level 2 (65.8%), and Level 3 (61.3%) tasks.

It is not readily apparent what the 40–70% range in the TSRs signifies, because there were several other variables involved in this result, such as the language levels of individual users, task types, DTs, task levels, topics, and (or) task directions. It could be postulated that low TSRs may not necessarily be due to the

participants' L2 proficiency level, but rather the design of the DTs in some tasks or insufficient direction may have been responsible for such low rates. For example, the TSR for the volunteer interview task (Level 2) was only 5.9% (1 success out of 17 conversation sessions); thus, in this case, task design (the DT or task types) may have been responsible for this low TSR. Additionally, some of the other tasks with lower TSRs, such as cancelling tickets (40.7%, 20 successes out of 42 sessions, Level 3), taking menu orders (37.5%, 39 successes out of 104 sessions, Level 3), giving advice on playing the guitar (33.3%, 2 out of 6 sessions, Level 2), appeared to have been particularly challenging. Further data collection (e.g., interviewing the participants) and qualitative analysis of the dialogues are needed to understand the low TSRs in relation to these tasks.

Table 4

Task Success Rates by Each Task Type and by Each Level in Task Chat

		Task Success Rate (%)	Valid Total Conversation Sessions	Successful Conversation Sessions
By Task Types	Personal Problems	60.7	56	34
	Missions	67.4	720	485
	Query/Questions	43.8	48	21
By Levels	Level 1 (*n* = 18)	69.9	259	181
	Level 2 (*n* = 24)	65.8	281	185
	Level 3 (*n* = 6)	61.3	284	174
Overall (Average / Total)		65.5	824	540

Conclusion

This article aimed at presenting a model of an L2 learning chatbot, informed by previous literature on chatbots and CALL as well as current developments within AI technology. To this end, we sketched out the design principles and architecture of Ellie, an L2 chatbot specifically designed and developed for L2 learning purposes. The quality assessment results revealed that Ellie has the potential to become an effective language learning companion in providing valuable opportunities to learn and practice an L2. The participants were able to maintain a relatively lengthy conversation (mean: 9.3 CPS) in English and engage in L2 problem-solving tasks with a substantial amount of meaning negotiation. This type of speaking experience is rarely provided within regular English classes in EFL contexts due to class size and time restrictions within the curriculum. We would like to highlight that Ellie could compensate for such limitations by serving as a conversational partner for L2 learners.

Despite these positive findings, we also noticed some limitations when applying Ellie within actual English learning classrooms during the pilot phase. First, Ellie occasionally failed to comprehend the participants' utterances due to noise in the classroom. This caused invalid conversation sessions, which could demotivate the users or even lower their confidence in speaking. Hopefully, advances in sound technology will soon resolve this practical issue. Second, there were several occasions in which Ellie could not respond adequately to the participants' utterances. This may have happened because the participants' utterances were not predefined in the API layer, or because of insufficient training. Either way, a larger amount of user data and machine learning appear to be essential in overcoming this type of limitation. Third, the DT structures in each task need to be further revised to enable users to perform the tasks as successfully as these tasks were designed to allow. Furthermore, continuous efforts to extending topic trees through follow-

up phrasing is required to sustain open conversation in the General Chat mode. Qualitative analysis of the data logs as well as the employment of statistical parameters, such as CPS or TSR, are expected to contribute to improvements in the quality of chatbot-based tasks. Nevertheless, we believe that Ellie could serve as a useful reference for future L2 chatbot projects, in terms of design and implementation of more effective chatbots.

Acknowledgements

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

References

Abdul-Kader, S. A., & Woods, J. (2015). Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications, 6*(7), 72–80. http://dx.doi.org/10.14569/IJACSA.2015.060712

Alm, A., & Nkomo, L. M. (2020). Chatbot experiences of informal language learners: A sentiment analysis. *International Journal of Computer-Assisted Language Learning and Teaching, 10*(4), 51–65. http://dx.doi.org/10.4018/IJCALLT.2020100104

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. http://dx.doi.org/10.1145/3442188.3445922

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, 32*(8), 827–877. https://doi.org/10.1080/09588221.2018.1535508

Bradeško, L., & Mladenić, D. (2012). A survey of chatbot systems through a Loebner prize competition. In *Proceedings of Slovenian Language Technologies Society Eighth Conference of Language Technologies* (pp. 34–37). Ljubljana: Institut Jožef Stefan.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint*. https://arxiv.org/abs/2005.14165

Chapelle, C. (1998). Multimedia CALL: Lessons to be learned from research on instructed SLA. *Language Learning & Technology, 2*(1), 22–34. https://www.lltjournal.org/item/2260

Chapelle, C. (2001). *Computer applications in second language acquisition*. Cambridge University Press.

Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter, 19*(2), 25–35. https://doi.org/10.1145/3166054.3166058

Coniam, D. (2004). Using language engineering programs to raise awareness of future CALL potential. *Computer Assisted Language Learning, 17*(2), 149–175. https://doi.org/10.1080/0958822042000334226

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering, 22*(5), 811–817. https://doi.org/10.1017/S1351324916000243

Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2020). A survey of deep learning and its applications: A new paradigm to machine learning. *Archives of Computational Methods in Engineering, 27*, 1071–1092. https://doi.org/10.1007/s11831-019-09344-w

Di Prospero, A., Norouzi, N., Fokaefs, M., & Litoiu, M. (2017). Chatbots as assistants: An architectural framework. In *Proceedings of the 27th Annual International Conference on Computer Science and Software Engineering* (pp. 76–86). IBM Corp.

Diederich, S., Brendel, A. B., & Kolbe, L. M. (2019, February). *On conversational agents in information systems research: Analyzing the past to guide future work* [Paper presentation]. 14th International Conference on Business Informatics, Siegen, Germany. https://aisel.aisnet.org/wi2019/track13/papers/1/

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Fedus, W., Zoph, B., & Shazeer, N. (2021). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint*. https://arxiv.org/abs/2101.03961

Fryer, L. K., Coniam, D., Carpenter, R., & Lăpușneanu, D. (2020). Bots for language learning now: Current and future directions. *Language Learning & Technology, 24*(2), 8–22. http://hdl.handle.net/10125/44719

Fryer, L. K., Nakao, K., & Thompson, A. (2019). Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in Human Behavior, 93*, 279–289. https://doi.org/10.1016/j.chb.2018.12.023

Godwin-Jones, R. (2021). Big data and language learning: Opportunities and challenges. *Language Learning & Technology, 25*(1), 4–19. http://hdl.handle.net/10125/44747

González-Lloret, M. (2003). Designing task-based CALL to promote interaction: En busca de esmeraldas. *Language Learning & Technology, 7*(1), 86–104. https://www.lltjournal.org/item/2419

Greyling, C. (2019, August 14). Chatbots: From unstructured data to conversation: Have a conversation with your customer. *Medium*. https://medium.com/@CobusGreyling/chatbots-from-unstructured-data-to-conversation-4bef5b014c47

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, J.-X., Lee, K.-S., Kwon, O.-W., & Kim, Y.-K. (2017). A chatbot for a dialogue-based second language learning system. In K. Borthwick, L. Bradley & S. Thouësny (Eds.), *CALL in a climate of change: Adapting to turbulent global conditions–short papers from EUROCALL 2017* (pp. 151–156). Research-publishing.net. http://doi.org/10.14705/rpnet.2017.eurocall2017.705

Joo, H. W. (2008). *A corpus-based analysis of vocabulary in the BEWL and the CSAT* [Unpublished master's thesis]. Korea University.

Khan, R. (2017). Standardized architecture for conversational agents a.k.a. chatbots. *International Journal of Computer Trends and Technology, 50*(2), 114–121. https://doi.org/10.14445/22312803/IJCTT-V50P120

Kim, H. (2018). *Designing L2 interactive tasks with an artificial intelligence robot* [Paper presentation]. Asia TEFL International Conference, Macau, China. https://www.slideshare.net/heyoungkim/designing-l2-interactive-tasks-with-an-artificial-intelligence-robot

Kojouharov, S. (2016, September 18). Ultimate guide to leveraging NLP & machine learning for your chatbot. *Chatbots Life*. https://chatbotslife.com/ultimate-guide-to-leveraging-nlp-machine-learning-for-you-chatbot-531ff2dd870c

Lee, J. H., Yang, H., Shin, D., & Kim, H. (2020). Chatbots. *ELT Journal, 74*(3), 338–344. https://doi.org/10.1093/elt/ccaa035

Lobo, J. (2017, October 23). What is a decision tree and why should my chatbot use it? *Inbenta*. https://www.inbenta.com/en/blog/decision-tree-chatbot/

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins.

Nunan, D. (2004). *Task-based language teaching*. Cambridge University Press.

Ó Broin, U. (2017, October 14). Personality brings life to chatbot user experience: Making chatbot interactions come alive. *Chatbots Magazine*. https://chatbotsmagazine.com/avoiding-a-clash-of-personalities-chatbot-design-is-no-different-3f0bcd30defd

Paulston, C. B. (1970). Structural pattern drills: A classification. *Foreign Language Annals, 4*(2), 187–193. https://doi.org/10.1111/j.1944-9720.1970.tb02033.x

Shah, A., Jain, B., Agrawal, B., Jain, S., & Shim, S. (2018). *Problem solving chatbot for data structures* [Paper presentation]. 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV. https://doi.org/10.1109/CCWC.2018.8301734

Shum, H.-Y., He, X.-D., & Li, D. (2018). From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), 10–26. https://doi.org/10.1631/FITEE.1700826

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Turing, A. M. (1950). I.—Computing machinery and intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Wang, Y. F., & Petrina, S. (2013). Using learning analytics to understand the design of an intelligent language tutor–Chatbot Lucy. *International Journal of Advanced Computer Science and Applications, 4*(11), 124–131. https://doi.org/10.14569/IJACSA.2013.041117

Willis, D., & Willis, J. (2001). Task-based language learning. In R. Carter & D. Nunan (Eds.), *The Cambridge guide to teaching English to speakers of other languages* (pp. 173–179). Cambridge University Press.

Yang, H., Kim, H., Lee, J. H., & Shin, D. (in press). Implementation of an AI chatbot as an English conversation partner in EFL speaking classes. *ReCALL*.

Zeroual, I., & Lakhouaja, A. (2018). Data science in light of natural language processing: An overview. *Procedia Computer Science*, *127*, 82–91. https://doi.org/10.1016/j.procs.2018.01.101

Zhou, L., Gao, J., Li, D., & Shum, H.-Y. (2020). The design and implementation of Xiaoice, an empathetic social chatbot. *Computational Linguistics*, *46*(1), 53–93. https://doi.org/10.1162/coli_a_00368

About the Authors

Heyoung Kim received her PhD in second and foreign language education from the State University of New York at Buffalo and is currently a professor at Chung-Ang University, South Korea. Her research interests include CALL, digital literacy, and task-based learning.

E-mail: englishnet@cau.ac.kr

Hyejin Yang received her PhD in applied linguistics and technology from Iowa State University and is currently a full-time researcher at Chung-Ang University, South Korea. Her research interests include CALL and language testing.

E-mail: hjyang1112@gmail.com

Dongkwang Shin received his PhD in applied linguistics from Victoria University of Wellington and is currently an associate professor at Gwangju National University of Education, South Korea. His research interests include corpus linguistics, CALL, and AI-based language learning.

E-mail: sdhera@gmail.com

Jang Ho Lee received his DPhil in education from the University of Oxford, and is presently an associate professor at Chung-Ang University, South Korea. His areas of interest are CALL, L1 use in L2 teaching, and vocabulary acquisition. All correspondence regarding this publication should be addressed to him.

E-mail: jangholee@cau.ac.kr