This is an **Accepted Manuscript** of a book chapter published as:
Bibauw, S., François, T., & Desmet, P. (2022). Dialogue systems for language learning: Chatbots and beyond. In N. Ziegler & M. González-Lloret (Eds.), *The Routledge Handbook of Second Language Acquisition and Technology* (pp. 121–134). Routledge. https://doi.org/10.4324/9781351117586-12

# Dialogue Systems for Language Learning: Chatbots and beyond

Serge Bibauw, Thomas François, and Piet Desmet

**Abstract**

Dialogue systems allow a user to interact, orally or in writing, with an automated interlocutor, whether it is referred to as a chatbot, a robot, a conversational agent or an intelligent personal assistant. We discuss the different typologies of dialogue-based computer-assisted language learning (CALL), the natural language processing (NLP) technology operating those systems, and the issues of their instructional design. We review the scientific findings on the cognitive, behavioral and emotional effects of dialogue systems. Finally, we provide recommendations for the use of dialogue-based CALL in foreign language learning and teaching, as well as for the development of new conversational applications.

## Introduction

Research in second language acquisition has long stressed the importance of meaningful practice in the target language (Gass & Mackey, 2015). However, foreign language learners are offered very unequal opportunities for interaction, especially spoken interaction. This situation has led computer-assisted language learning (CALL) researchers since the 1980s to investigate the use of computers as tutors or language partners, calling these systems either *chatbots*, *dialogue systems*, *conversational agents*, *robots*, or *intelligent personal assistants*.

In this chapter, we use the term *dialogue systems for language learning* as an umbrella term, synonymous with *dialogue-based CALL*, to encompass all technology allowing a human to converse with an automated agent, i.e., a virtual or physical robot, in the context of foreign or second language (L2) learning. The systems have been designated by different terms, and they vary in modality (written, spoken, or multimodal), in interface (a computer, a smartphone, a smart speaker, or an actual robot), in interactional context (an isolated dialogue as an exercise, an agent built as a social partner or as a tutor, a non-player character in a video game), and in dialogue-management technology (from keyword recognition to machine learning). Yet, all these systems constitute a single type of affordance of technology for language learning (Bibauw et al., 2015, 2019).

Bibauw et al. (2019) have proposed an operational definition grounded in the literature for the concept: dialogue-based CALL, or dialogue systems for language learning, includes any system or application where a dialogue, i.e., various conversational turns, held with virtual agents controlled by the computer, is used as an L2 learning task. Virtual agents set it aside from computer-mediated communication (CMC), where the interlocutor is another human. The multiple turns distinguish it from item-based tutorial CALL, organized around isolated items rather than as a conversation. And the fact that the dialogue *is* the task sets it apart from pedagogical agents, which provide scaffolding, sometimes in the learner's first language (L1), on an external, non-dialogic task.

The use of dialogue systems for language learning is based on the shared assumptions that meaning-focused interactional practice of the target language is essential to SLA (Gass & Mackey, 2015) and that, while human conversations remain the gold standard and the end-goal, dialogue systems

can provide valuable opportunities for such practice to learners—particularly in foreign language learning contexts—who rarely have opportunities to interact in the target language (Sydorenko et al., 2019). Over human interlocutors, they even present the advantages of permanent availability, equanimity in case of repetitions or corrections, and a lower-anxiety environment, which might help beginning or apprehensive learners (Alemi et al., 2015). From an interactionist perspective of SLA, dialogue systems offer input, output, and interaction, including interactional feedback on one's messages. Many systems also provide corrective feedback and negotiation of meaning, which are known to have a major impact on language learning (e.g., Sato & Lyster, 2012). As a technology, dialogue systems fulfill many characteristics of what Thomas et al. (2012) called 'contemporary CALL': they provide feedback, 'mirror authentic environments', and 'provide users with opportunities for greater control, creativity, and freedom' (pp. 8-9).

## Historical Perspectives

In the last forty years, efforts towards simulated L2 conversations have emerged in different domains, under different names, and with different focal points. We present a brief historical overview of the various types of dialogue systems for language learning.

*ICALL and Intelligent Tutoring Systems.* The first attempts of dialogue-based CALL date back to the early 1980s, in intelligent tutoring systems. Aware of the need to go beyond grammar practice in isolated items, "intelligent CALL" (ICALL) researchers started to leverage Natural Language Processing (NLP) techniques to allow for freer meaning-focused interactions (see Heift & Schulze, 2015 for a timeline of tutorial CALL). Various systems used very limited environments (*microworlds*) or sets of possible commands—inspired by text adventure games—to reduce the unpredictability of the user input. Most efforts were dedicated to the language parsing issues and the provision of automatic corrective feedback on open-ended input (e.g., Holland et al., 1995).

*Computer-Assisted Pronunciation Training (CAPT).* While almost all early ICALL systems were text-based, other research groups were working exclusively on speech. At the end of the 1990s, they capitalized on the advances in signal processing and automatic speech recognition (ASR) to develop speech-based systems focusing on correcting pronunciation (see Eskenazi, 2009). Some integrated the pronunciation correction in meaningful—though heavily constrained—dialogues, but naturally, the focus was essentially on pronunciation assessment (e.g., Engwall, 2012).

*Spoken Dialogue Systems (SDS).* In the 2000s, teams that had developed spoken dialogue systems for informational and commercial purposes on the telephone started to explore language learning applications (Eskenazi, 2009). With a stronger NLP background, they brought more complex dialogue management strategies, allowing for freer conversations, organized around specific tasks (ordering in a restaurant, booking a flight…). Some of these systems integrate the dialogues in 3D virtual worlds, with embodied—i.e., portrayed by an avatar—conversational agents and multimodal interfaces, as in the dialogue-based game presented in **Figure x.1**. In contrast, they do not always implement corrective feedback or any type of focus on form, emphasizing rather the output opportunities and the interactional feedback as their strong points.

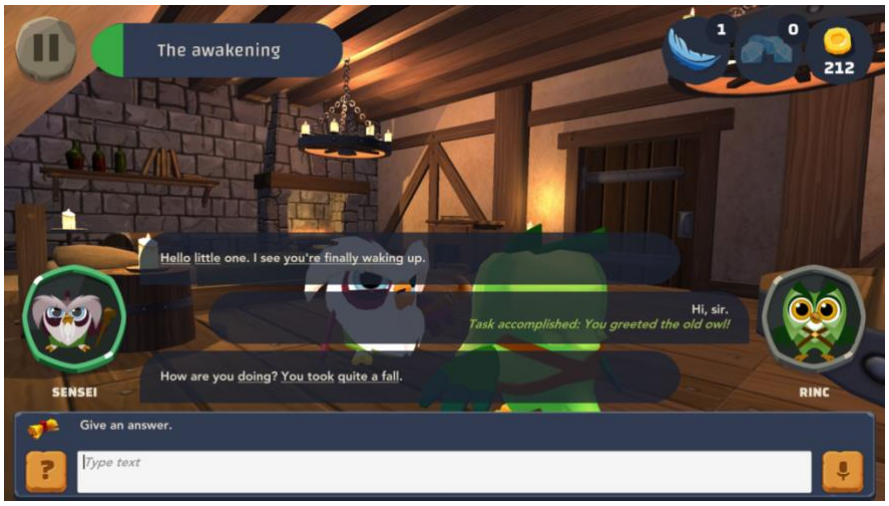


**Figure x.1** *Language Hero* is an example of a dialogue system for language learning integrated inside a video game.

*Social robots.* Recent developments have occurred in Human-Robot Interaction, using physical robots for the same purpose of maintaining a meaningful conversation (see van den Berghe et al., 2019 for a review). The dialogue technology behind it is often similar to SDS, hence sharing similar opportunities and challenges, but this type of research often appears in a separate field, under the label of Robot-Assisted Language Learning (Han, 2012). The physical presence of—usually—a single robot also means that it tends to be used in the classroom, in more collaborative ways, in comparison with more individual practice on computer-based SDS.

*Chatbots.* Almost none of the previous systems were ever usable by the general public, mostly remaining at the level of research prototypes. Chatbots, on the other hand, were created by computing enthusiasts to be openly accessible. Text-based, chatbots typically use relatively simple pattern-matching techniques to try to respond to the infinitely-unpredictable user messages. From the mid-2000s, certain chatbots were developed for language learning (e.g., the early Dave for ESL), and existing ones were used by language teachers (see Fryer et al., 2020). The main limitation of traditional chatbots is that, because they are open-ended and reactive, conversations are often aimless and incoherent, and the initial curiosity can quickly wear off. Some recent chatbot attempts by language learning applications (e.g., Duolingo Bots, introduced in 2016 but later retired, supposedly temporarily), however, could offer more pedagogically and interactionally sound conversations.

*Intelligent Personal Assistants (IPA).* Finally, the emergence of Intelligent Personal Assistants—such as Siri and Alexa—has recently opened new possibilities readily available for most learners. While IPAs are not designed for L2 learning, teachers and researchers have devised ways to exploit them interactionally with learners. With smartphones' IPAs, it will mostly consist of giving instructions to the learners to use the IPA to accomplish a certain task or dialogue result. With Alexa and similarly flexible systems, the instructional developer can program sets of commands and interactions that will lead to a specific exchange (Dizon, 2020).

## Critical Issues and Topics

Dialogue-based CALL systems are diverse, differing in interactional design, technological design, and instructional design. An important research question is hence to clarify the design and implementation choices for developers and practitioners, and the pedagogical consequences of these choices.

### *Types of Systems and Interactivity*

Dialogue systems are regularly categorized based on disciplinary or historical criteria (as above), or on interface criteria, such as modality: 'chatbots' are typically text-based, while 'dialogue systems' are in majority voice-based. The supporting device typically plays a major role in how laypersons distinguish systems: many users see Siri and Alexa as different technologies because they operate primarily on different devices (smartphones vs. smart speakers), even though the NLP techniques that make them work can be identical; a similar reasoning applies to the distinction between telephone-based dialogue systems and robots.

From a technological point of view, Jokinen and McTear (2010) distinguish two types of spoken dialogue systems: *task-oriented systems*, which use the dialogue to accomplish a task, such as booking a flight or planning a vacation, and *non-task-oriented systems*, also called *open-ended systems*, where the conversation has no specific pre-established goal. Task-oriented systems are further categorized into *system-initiative* systems, where the user is restricted to short and predictable answers, and *conversational* systems, which leverage more complex processing to allow for freer interaction, closer to spontaneous speech.

From a language learning perspective, however, any taxonomy of systems should be grounded in the different affordances they offer for language learning. In their research synthesis of dialogue-based CALL, Bibauw et al. (2019) proposed a typology of systems based on the constraints each system imposes on the form and meaning of the user's input, considering that these constraints will directly influence the learning activity and its focal point. This typology distinguishes (1) *narrative systems*, such as branching dialogues where the user chooses among a list of pre-set utterances, (2) *form-focused systems*, where the meaning is imposed (e.g., by pre-establishing part of the utterance or providing the expected meaning) and all the focus is on form—pronunciation or target structures—through corrective feedback, (3) *goal-oriented systems*, which leverage task-oriented dialogue systems to allow for less constrained input by setting a context and a task to accomplish, and (4) *reactive systems*, such as classical chatbots, which offer open-ended interactions by leaving the initiative to the user and trying to provide a relevant response to each user's message (see Bibauw et al., 2019 for analysis of cases and discussion of implications).

Constraints on user production are directly linked to the degree of interactivity that each system offers. Drawing from McMillan's (2002) model of interactivity for web-based communication, it is possible to devise a model of interactivity for dialogue systems designed across two dimensions: (a) the *direction of adaptivity*, whether there is a continuous, reciprocal adaptivity to the interlocutor's responses (*two-way* adaptivity: the interactant who has the initiative adapts back its following turn to their interlocutor's action) or only an intermittent one, from one side only (*one-way*: the responder has a reactive role only), and (b) the *level of user control*, mainly defined by the initiative of the dialogue—user control is low when the system holds the initiative, and high otherwise. The four quadrants delimited on this bi-dimensional space, depicted in **Figure x.2**,

correspond to four types of interactivity, from *system-guided* (one-way adaptivity, low user control) to *full-interactivity* (two-way, high user control), where the four types of dialogue-based CALL systems identified in Bibauw et al. (2019) fit perfectly.

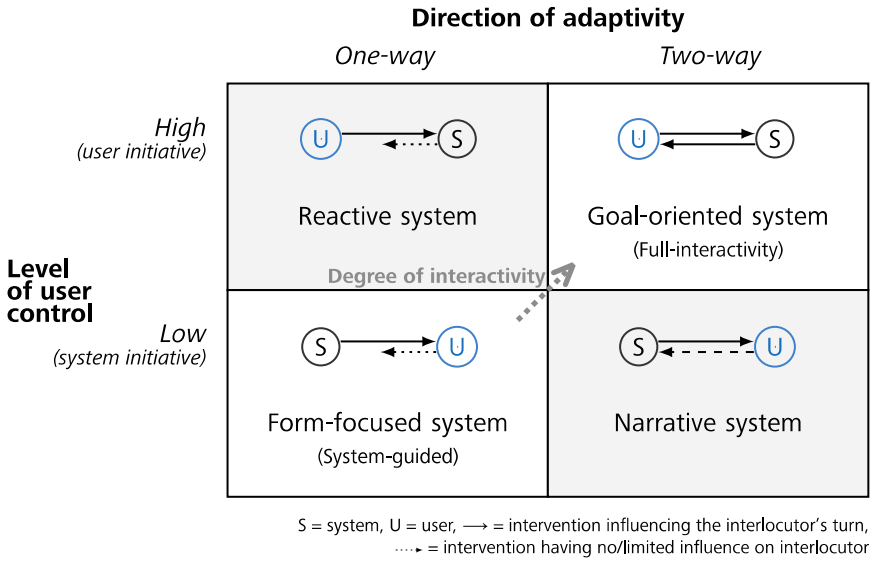


**Figure x.2** Model of interactivity in dialogue systems for language learning (based on McMillan, 2002).

It is particularly the system-guided (form-focused systems) and full-interactivity (goal-oriented systems) options that offer the most promising affordances for language learning. The system-guided applications, in which the virtual agent does not (need to) take into consideration the user's replies to its questions or injunctions (for instance, it could mimic an interviewer following their list of questions, without reacting upon the answers), offers a natural evolution of item-based tutoring systems: by focusing on corrective feedback rather than interactional flexibility, it emphasizes L2 accuracy and form-focused practice.

On the other hand, the full-interactivity type of system provides an ideal environment for the autonomous practice of spontaneous, synchronous interaction—either spoken or written—, in particular for fluency development. While the context is usually determined, the user is relatively free to lead the conversation, and the system seeks to react appropriately. This highly interactive form of dialogue is relatively close to a human-human conversation. It is permitted by a predefined task or goal, which allows the system to circumscribe the domain and the possible conversational paths without locking the range of possible actions. This type of dialogue satisfies the first four key features of technology-mediated task-based language teaching, as identified by González-Lloret and Ortega (2014)—primary focus on meaning, goal orientation, learner-centeredness (allowing the learner to use a variety of linguistic resources), and holism (related to real-world communicative activities)—and constitute as such an ideal technology for autonomous task-based practice. However, it is also technologically more challenging to develop, considering the complexity of dialogue management.

### *Dialogue Management Technology*

There is much diversity in the technology that operates dialogue systems, but it usually involves a series of complex NLP techniques. As this constitutes an entire field of NLP research (see Jurafsky & Martin, 2020, chap. 24), we can only skim over the question here.

The simplest approach to dialogue management is exhibited by rules-based chatbots—e.g., those programmed in AIML or Chatscript. Such chatbots try to respond to any possible user message by building large knowledge bases of pattern-matching rules (e.g., '*my mother|mom*' ⇒ 'Tell me more about your mother.') (Jurafsky & Martin, 2020). The limitations of this reactive approach, however, rapidly become evident. These systems allow for a few conversational turns, often driven by playful attempts at tricking the bot and exposing its shortcomings, but curiosity rapidly wanes. With such reactive systems, the lack of an interactional goal and the absence of conversational memory make it almost impossible to build up a longer exchange or to reach a conclusion to the conversation.

Very differently, task-oriented dialogue systems developed by NLP researchers traditionally implement a pipeline of modules: automatic speech recognition (for spoken input), natural language understanding, dialogue management, and natural language generation, plus speech synthesis if necessary (see Chen et al., 2017). This means that, after the vocal signal has been transcribed into text (e.g., 'Can I have a plastic bag, please?'), the message is first interpreted (natural language understanding)—for instance, via an intent classifier, to recognize the dialogue act intended (here, e.g., request-item), and an entity recognizer, which will extract isolated pieces of information (*item* = plastic bag)—, then integrated into the system's representation of the dialogue (dialogue state tracking: *user* = customer, *current step* = paying, etc.) to decide what following action to undertake (dialogue policy: *next act* = accept + give-item), and then convert this semantic action into a message (natural language generation: 'Of course. Here you are').

Each of these modules is an NLP challenge in itself and multiple strategies exist for each, which can generally be categorized as either rule-based or corpus-based (Jurafsky & Martin, 2020). In most commercial and research systems, natural language understanding is nowadays realized via data-driven machine learning techniques: intent and entity recognizers can be trained on a large annotated corpus of interactions (Serban et al., 2018). Dialogue state tracking and dialogue policy can also be learned from dialogue corpora, but in most commercial applications, they are built as handcrafted—i.e., created by a human—branching rules for the specific task at hand, eventually with probabilities of transitions between the branches learned from previous interactions (*probabilistic rules*). Natural language generation—the transformation of the semantic representation into an utterance—can also be learned from data, either in retrieval (the model selects the most appropriate utterance from an existing corpus of responses) or fully generative fashion (the model learns—with deep sequential learning techniques on big data sets—to recreate new sentences). However, again, many commercial applications rather use a knowledge base of handcrafted sentences, which trade the flexibility of data-driven approaches for higher reliability.

As one can see, research and commercial dialogue systems often follow radically divergent strategies. This phenomenon has been intensified by the rise of deep learning in NLP, with research on dialogue systems now essentially focused on using large corpora and complex neural networks, and even attempting end-to-end solutions—i.e., learning to select or generate an appropriate response directly from user input, requiring even larger corpora—rather than the traditional

pipeline approach we mentioned above, where the input passes through successive independent modules (Serban et al., 2018). Meanwhile, production systems continue to rely on hand-built rules and ad-hoc implementations, because of their accurate and predictable results (Chen et al., 2017). Recent commercial 'conversational AI' platforms, such as Google Dialogflow, Microsoft Bot Framework, IBM Watson Assistant, or the open-source framework Rasa, which together power most enterprise virtual agents, tend to combine data-driven understanding (intent and entity extraction) with handcrafted decision trees for dialogue state and policy, and prewritten messages for natural language generation.

Dialogue systems designed for language learning vary widely, depending on their research or applicative purpose, but most use hybrid approaches and do not hesitate to resort to ad-hoc solutions (Bibauw et al., 2019). Clever use of specific interface and task features can circumvent the limitations of a specific technology. For instance, Duolingo Bots' behavior was entirely handcrafted, thus limiting their ability to process unforeseen input, but constrained the virtual keyboard on smartphones and tablets to only accept text for which the bot knew how to respond. Intelligent task design and more controlled domains, by reducing the complexity and the unpredictability of the conversation, offer promising solutions for dialogue-based CALL.

### *Instructional Design and Tasks*

The complexity of dialogue management could give developers the impression that dialogue-based CALL is only an NLP problem. In the early literature, publications were mostly discussing technological challenges. However, the design of the interactional context, the task at hand, and the conversational content have equally important pedagogical consequences. Firstly, putting learners in front of a conversational agent without a precise task or instruction would be equally ineffective as telling two learners only to 'speak together' without a topic or a purpose. Secondly, learning only occurs from adequately designed interactions: 'children do not just learn by being exposed to a tutoring robot' (Belpaeme et al., 2018, p. 337). Thirdly, there is a tremendous technological distance between a system able to conduct specific tasks and one able to discuss any topic. Usable and satisfactory open-domain dialogue systems are still far off. Thus, even for solving technological shortcomings, instructional design is key, and discussing it, as well as disclosing instructional choices in research publications, of critical importance.

An instructional design framework of dialogue-based CALL necessarily starts from a definition of learning goals. Previous dialogue systems have essentially targeted either linguistic form, focusing on specific L2 structures (e.g., German dative in Wilske, 2015) or on pronunciation training (e.g., Engwall, 2012), or skills development (e.g., Dizon, 2020).

Deciding on a certain type of learning goal should orient the developer toward a certain type of dialogue system: narrative systems, close to interactive fiction and adventure games, offer mostly input and are thus ideal for primarily reading or listening practice; system-guided, form-focused systems are the go-to type for accuracy, grammar, and pronunciation training; and fully-interactive goal-oriented systems offer a lot of potential for fluency development in productive skills.

Beyond the type of system, multiple instructional features have to be decided in relation to the learning goals. Will the interface be oral-only, written-only, or multimodal? Does it require a visual embodiment of the agent—an avatar—with facial expressions or gestures? An important design choice is also the instructional focus: dialogue-based CALL lends itself well to both *focus-on-*

*meaning* and *focus-on-form*. In the latter case, the provision of corrective feedback should be strongly considered, either in a communicatively integrated way, possibly as recasts (e.g., Wilske, 2015), or in more explicit and externalized ways (e.g., Cornillie et al., 2013).

The design of the conversational tasks themselves is critical as well, and closely linked to the system choices (see Sydorenko et al., 2019). Will the application offer one or multiple tasks? In the case of multiple tasks, what will be the progression? Is the expected dialogue symmetrical—with both interlocutors sharing the initiative—or asymmetrical—where one interlocutor leads the whole exchange? In the latter case, will it be system-guided or user-guided? How should the task and its successful completion be internally modeled in the system? What will be the content of the virtual agent's messages? Will modeling of the expected responses be offered, and if so, how will modeling be integrated? What type of scaffolding will be provided to the learner? Will the application adapt the dialogue to the learner's proficiency level, for instance by implementing multiple versions of the task or the messages, varying in complexity?

It is also essential to inform the conversational tasks design from the rich framework offered by task-based language teaching research (González-Lloret & Ortega, 2014). It offers insights regarding the effects of the sequencing of the tasks, particularly in terms of task complexity (Ziegler, 2016). For instance, adjusting the task complexity along resource-directing factors, such as reducing the number of elements mentioned in a dialogue or referring to 'here and now' rather than to 'there and then', might allow learner to dedicate more attention to developing the language of their responses (Robinson & Gilabert, 2007). Dialogue systems also open the possibility to repeat a conversational task several times with slight adjustments of both content—for motivation—and task complexity—for learning optimization.

In summary, there are many different ways to categorize dialogue systems for language learning, depending on their modality, their goal-orientation, or their degree of interactivity, adaptivity and user control. There is also a wide range of NLP options to carry out these dialogues, but in general, dialogue systems for language learning tend to use ad-hoc and simpler approaches, and pay a lot of attention to the instructional design, including the adequation to the learning goals, the provision of scaffolding and feedback, and the design of conversational tasks.

## Current Contributions and Research

In the last fifteen years, empirical research on dialogue systems for language learning has increased steadily. The budding field has now a decent amount of studies, offering emergent evidence of the effects of dialogue-based CALL in terms of learning outcomes and learners' attitudes, as well as the relative effects of specific design choices.

### *Effects on Learning Outcomes*

In the first meta-analysis on the subject, Bibauw et al. (2022) identified an overall medium effect of dialogue systems on language learning outcomes ($d$ = 0.59). This is slightly smaller than the observed effect of interaction with native and non-native speakers ($d$ = 0.75) (Mackey & Goo, 2007), which is coherent with a vision of dialogue systems as a suboptimal stand-in for human interlocutors.

More precisely, we know that dialogue systems lead to significant learning gains in vocabulary and grammatical outcomes, as measured in knowledge tests, and in general proficiency and accuracy, when measured in productive tests. How much they might affect productive complexity, and particularly fluency, is still unclear, due to a lack of evidence. It might in particular be necessary to implement long-term intervention studies to be able to observe significant effects on fluency and complexity, but the difficulty of content development has until now reduced most dialogue systems to short-term interventions. Effects were similar for speech-based and text-based systems, and they were equivalent across spoken and written test modalities, but they grew stronger when system modality and testing modality matched, indicating that, while transfer does happen across modalities—e.g., practicing orally also develops writing skills—, it is only partial. The sustainability of the learning gains also seemed high, as the decline between immediate and delayed post-tests was insignificant (Bibauw et al., 2022).

Recent studies are also providing rich insights and confirmation of the potential for dialogue systems in facilitating L2 learning and development (e.g., Hong et al., 2020). Dizon (2020) conducted a 10-weeks in-classroom intervention with Alexa, in which students interacted individually and freely with a smart speaker for 12 minutes each week, inspired by a list of commands they could use, including 'word of the day' quizzes, interactive storytelling skills like Earplay and socialbots interaction skills. The study evidenced significant learning gains in speaking proficiency—although not in listening comprehension—in comparison with the control group, confirming the usefulness of intelligent personal assistants for offering speaking practice in foreign language contexts.

Reviewing the related field of robot-assisted language learning, van den Berghe et al. (2019) identified mixed, but generally positive effects on vocabulary learning. Smaller learning gains were observed in multiple-sessions studies with young children, which could be explained by insufficient time-on-task. Yet, in single-session studies and studies with older children and adults, the learning gains from interacting with the robot were larger. When compared with a human tutor or peer, robots did not outperform interlocutors, but they demonstrated they can reach similar effects. (Note however that, in a few studies, robots were teleoperated by a human rather than programmed to work autonomously, as a dialogue system. This logically leads to behaviors closer to human reference.)

A domain that has seen a notable amount of dialogue-based CALL studies is pragmatics instruction. Using in some cases relatively accessible technological tools, such as page-by-page interaction with pre-recorded messages or self-analysis of their speech by the learners, researchers have been able to demonstrate that simulated conversations had strong effects on the acquisition of pragmatically appropriate speech acts (Alemi & Haeri, 2020; Sydorenko et al., 2018) and of formulaic expressions (Taguchi et al., 2017). Those studies also attest to the merit of embodied agents, either through a physical robot or presented dynamically on-screen (video recording or avatar), for the acquisition of politeness, gestures, and pragmatics in general (Alemi & Haeri, 2020).

### *Behavioral and Emotional Effects*

Besides learning outcomes, dialogue systems are also used for their potential to positively affect learner's attitudes. Multiple studies have shown that learners felt less anxious when interacting with a dialogue system than with a human interlocutor (e.g., Alemi et al., 2015). A neurological

study on brainwaves observed that students who were discussing with a chatbot were in a more relaxed state than those who were talking to a human interlocutor, although their attention was lower than those interacting face-to-face with their peer (Hsu, 2020).

Dialogue systems have even post-treatment effects, with learners reporting more satisfaction and confidence, more interest and motivation, lower anxiety, and demonstrating more engagement after interactions with robots (Alemi et al., 2015; Hong et al., 2020; Lee et al., 2011). They can also affect positively learners' willingness-to-communicate, especially when implementing conversational strategies (including modified input) and affective backchannel confirmations (Ayedoun et al., 2019).

Yet, like any technology, part of the interest could be the result of a novelty effect, which could wear off and bring down learners' motivation (Sydorenko et al., 2019). It is thus essential to maintain engaging and varied conversations if longer time-on-task is hoped for. Van den Berghe et al. (2019) also warn that communication breakdowns and technical issues might accentuate this drop in interest over time, as current NLP techniques are still susceptible to recognition failures or inadequate response selection.

### *Differential Effects of Systems*

A couple of studies have compared the effectiveness of different versions of a certain dialogue system. Wilske (2015) compared a constrained form-focused system, based on gap-filling, with a goal-oriented system where the user input was free and either recasts or metalinguistic feedback were provided. While her findings are somewhat limited by reduced sample size, the form-focused system produced the strongest immediate results, while outcomes for the goal-oriented system with recasts seemed to last longer. The latter also induced stronger gains in speaking fluency, but only in one of the tasks.

Regarding physical robots, van den Berghe et al. (2019) conclude that the most promising results are obtained when a robot is used as a teaching assistant or as a peer, rather than as an independent tutor. They hypothesize that the current state of dialogue management technology might not be sophisticated enough for robots to provide effective tutoring on their own. This supports the emphasis in dialogue-based CALL on the use of virtual agents in conversational tasks, rather than to impart direct instruction.

A recent study by Engwall and Lopes (2020) comparing four robot behaviors with pairs of learners demonstrated that, overall, learners preferred an 'interviewer' robot, which would ask questions to each learner at a time, at least at an initial stage. Behaviors where the robot would either mostly keep the floor ('narrator') or, conversely, encourage learners to address each other, were less appreciated. However, in subsequent sessions and for more proficient learners, a more dynamic and personalized behavior ('interlocutor'), which targeted a three-way interaction between the robot and the peers, was preferred. This finding suggests the need for adaptive system behavior—both learner-adaptation and process-adaptation.

Bibauw et al.'s (2022) meta-analysis demonstrated significant effects of form-focused and goal-oriented systems, although there is insufficient empirical data to show an effect for narrative and reactive systems. Importantly, the addition of corrective feedback to any dialogue system almost doubles the observed effect ($d$ = 0.38 without, $d$ = 0.68 with implicit, and $d$ = 0.73 with explicit

feedback), confirming the well-established pedagogical importance of feedback, and the fact that corrective feedback makes a difference even on top of interactional feedback. A similar multiplier effect accompanies the gamification of the experience, which encourages the consideration of game-based motivational elements in future systems.

## Recommendations for Practice and Development

### *Dialogue Systems in Instructed SLA*

Teachers identify two priority uses for dialogue-based CALL: as extra out-of-class conversational practice, and as diagnostic tools for formative assessment (Timpe-Laughlin et al., 2020).

The first obstacle facing anyone willing to use dialogue systems in a language learning context is the availability of applications. Among the very few publicly accessible systems designed for SLA are the Mondly chatbot[1] and the virtual reality-enabled ImmerseMe[2], both available for many target languages and offering system-guided interactions focused on pronunciation. English language learners can also practice in open-ended written conversations with relatively limited chatbots such as Andy[3] and Tutor Mike[4].

Reactive systems, such as general-purpose chatbots and—to a lesser extent—intelligent personal assistants (e.g., Google Assistant), might not offer the richest affordances for language learning, due to the limited depth of conversations and their relative aimlessness in the absence of tasks, but they have the major advantage of being easily and freely available. Amazon's Alexa has in particular been used in various educational settings as the versatility offered by external 'skills' allows for extensibility of the platform, for instance with interactive storytelling or word-learning content (Dizon, 2020).

For language teachers who would like to use such systems with their learners, the key will be to design a series of tasks and instructions to guide the learners in their interactions. Tasks should align with the learning goals, and encourage conversational exploration. Make sure however to pilot them with the target system, to ensure that the tasks can be achieved with it as foreseen and to identify possible sources of communication breakdown for learners.

The success of dialogue-based CALL with learners will strongly depend on the preparation of the learners before the interaction with the automated agent, and on the scaffolding provided, either by the agent or the teacher (Timpe-Laughlin et al., 2020). In particular, the place of modeling should be anticipated (Sydorenko et al., 2019): will the system model the speech acts later expected from the learners or does it require prior learning of certain language resources? Integrating modeling in the virtual dialogue enables the system to be used autonomously and ensures a smoother progression from exposure to use, but it is not always feasible, especially if the teacher has no

[1] https://www.mondly.com/

[2] https://www.immerseme.co/

[3] https://andychatbot.com/

[4] https://www.rong-chang.com/tutor_mike.htm

control over the dialogue agent. In such cases, the pre-task phase will need to prepare the learners and provide models before they start the conversation.

Finally, dialogue systems and robots should not be seen as strictly individual learning tools. On the contrary, various studies have demonstrated that collaborative use of dialogue-based CALL applications not only has no negative effect (Dizon, 2020) but can be very fruitful (Engwall & Lopes 2020). In particular, having learners interact in pairs with a dialogue system allows for circumventing both learners' language deficiencies and the system's limitations, by turning to the peer in case of communication breakdowns.

### *Development of Dialogue Systems*

Developing a new dialogue-based CALL application is a complex and long-term endeavor. It is not, however, reserved for NLP programmers, as certain tools make the creation of a less ambitious conversational system more accessible. 'Conversational AI' cloud-based commercial platforms, such as Google DialogFlow[5], Facebook Messenger Platform[6] , or Microsoft Bot Framework[7], make it feasible to program a chatbot without any coding or server management. They offer a probabilistic natural language understanding engine, composed of intent recognition and entity extraction, combined with a decision tree for response selection, in other words, a relatively advanced set-up, much preferable to classic pattern-matching chatbots. To avoid being confined to a proprietary framework, open-source frameworks such as Rasa[8] and DeepPavlov[9] offer equally powerful tools, but will require some familiarity with Python programming. The main limitation of all these solutions is that they are primarily designed for question-answering enterprise chatbots, hence trying to respond to the user's needs in as few turns as possible. Of course, in L2 practice, longer conversations are desirable, and adjustments will be necessary.

Designing a dialogue system starts from the exploration of human dialogues through corpora, or through wizard-of-oz experiments, in which users' messages are actually answered by a human writing the responses to be uttered by the system. The design process should be iterative, going through multiple rounds of prototyping, testing and incremental development. Awareness of technological shortcomings and potential issues is also essential: the state-of-the-art in the multiple techniques required to process dialogue, while rapidly advancing, is still error-prone (Belpaeme et al., 2018).

Developers should also pay a lot of attention to the tasks and the interaction design (and report those choices in detail in publications). Systems should try to offer opportunities for longer and increasingly complex conversations, by incorporating more negotiation and growing complexity in the tasks. For instance, in a dialogue about scheduling a meeting, the virtual interlocutor could have limited availability or suddenly remember another engagement, requiring more back-and-forth proposals (Timpe-Laughlin et al., 2020). To maintain motivation and engagement, they

---

[5] https://cloud.google.com/dialogflow

[6] https://developers.facebook.com/docs/messenger-platform/built-in-nlp/

[7] https://dev.botframework.com/

[8] https://rasa.com/

[9] https://deeppavlov.ai/

should also avoid repetitiveness, which requires a least a minimal conversational memory of both system and interlocutor's previous turns (Engwall & Lopes, 2020). Two other key factors of pedagogical success are scaffolding and feedback (Timpe-Laughlin et al., 2020). Implementation of corrective feedback can make the difference between negligible and significant impact (Bibauw et al., 2022). The quality of corrective and interactional feedback is crucial to ensure uptake and acquisition.

## Future Directions

*Development and access to dialogue systems for autonomous language learning*. The demand for online learning is higher than ever, and for language learning, this often requires new ways for the autonomous practice of spontaneous written or oral production. Dialogue systems can offer an ideal tool for meaning-focused interaction in this context, potentially less anxiety-ridden and more adaptable to each learner's proficiency level. However, this will require a very systematic effort to further develop existing prototypes of dialogue systems and to make them accessible to the general public. Strategic alliances between research-oriented designers and industry developers will probably be needed to take on this challenge, considering the amplitude of the required developments. Following the initiative of Timpe-Laughlin and colleagues (2020), it is also essential to work with language teachers on their expectations and needs for such systems, as the success of the technology will depend on its adoption by teachers and learners.

*Extended and comparative effectiveness studies*. We now have an emerging body of studies on the effectiveness of dialogue systems for language learning. Yet, we lack evidence of effects on multidimensional constructs such as fluency and complexity. To be able to achieve visible learning gains in those areas, longer interventions are necessary, which require systems designed for longer time-on-task (most studies to date had interventions close to one hour). On another note, research should move beyond single-intervention studies and comparisons with no-system control conditions, toward comparisons between different implementations of simulated dialogues. More precise insights can be gained from the observation of alternating features in a specific system.

*Dialogue systems for SLA research.* The designed and focused interactions offered by dialogue systems constitute an ideal environment to research SLA theories, particularly to test interactionist hypotheses. They offer reproducible conditions of interactions, something that could be difficult to achieve with human interlocutors, while remaining realistic (Taguchi et al., 2017). Besides, abundant process data, including keystroke information, audio signal, and interface engagement, can easily and systematically be collected during the dialogue. This offers major potential for real-time analysis of learner's language and behavior in spontaneous practice conditions. By combining these technical advantages with varying features or characteristics of the system's behavior or implementation, possibly in a cross-sectional and randomized fashion for users, researchers can observe very precisely how the interactional variations influence learner's processes and performance.

*Dialogue systems for language testing.* Dialogue-based CALL also presents rich opportunities for the assessment of spontaneous L2 speaking or writing abilities (Litman et al., 2018). Due to their predictable and reproducible nature, simulated conversations can be used as an ideal computer-guided speaking test. Independently of the analysis of learners' performance (automatically or by a human rater), such interactions offer possibilities for observing generic productive proficiency,

target structures use, but also conversational strategies and dialogue fluency. They could be used both for allowing formative assessment on speaking and writing performance outside the classroom, as well as for summative assessment in higher stakes context.

*Adaptivity and personalization*. Finally, a very promising path for future dialogue systems' development is adaptivity. Most automated agents do not adapt their language to the proficiency level of their interlocutors, thus limiting their target audience to either very homogeneous or advanced groups of learners. Implementing adaptivity into their conversation would make them much more flexible and would provide scaffolding to learners, including by potentially offering modified input and other forms of negotiation of meaning when the learner signals a lack of understanding. With a precise detection and representation of the learner's interlanguage, the virtual agent could even optimize the pace of introduction of new vocabulary or target structures, or encourage the learner to use known-but-rarely-used words and structures. However, such adaptive behavior in dialogue is particularly complex to achieve, considering the already-complex dialogue management processes.

In general, dialogue systems offer very rich opportunities for language learners, testers, and researchers. While the technology still has a lot of ground to cover before reaching fully adaptive and natural interactions, it can already be put to use in carefully designed conversational tasks, with clear positive effects. The technology only needs more developers, researchers, and practitioners to bring it to its full potential.

## Further Readings

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

> Research synthesis of the field of dialogue systems for language learning, covering historical, conceptual, typological, methodological, and effectiveness aspects.

Belpaeme, T., Vogt, P., van den Berghe, R., Bergmann, K., Göksun, T., de Haas, M., Kanero, J., Kennedy, J., Küntay, A. C., Oudgenoeg-Paz, O., Papadopoulos, F., Schodde, T., Verhagen, J., Wallbridge, C. D., Willemsen, B., de Wit, J., Geçkin, V., Hoffmann, L., Kopp, S., … Pandey, A. K. (2018). Guidelines for designing social robots as second language tutors. *International Journal of Social Robotics*, *10*(3), 325–341. https://doi.org/10.1007/s12369-018-0467-6

> The paper summarizes approaches and insights specifically for developing social robots for language learning.

Jurafsky, D., & Martin, J. H. (2020). Dialogue systems and chatbots. In *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed. draft, pp. 487–517). https://web.stanford.edu/~jurafsky/slp3/26.pdf

> The chapter covering dialogue systems in the reference NLP textbook provides an up-to-date overview of the subject, from a technological point of view.

**References**


Alemi, M., & Haeri, N. (2020). Robot-assisted instruction of L2 pragmatics: Effects on young EFL learners' speech act performance. *Language Learning & Technology*, *24*(2), 86–103. https://doi.org/10125/44725

Alemi, M., Meghdari, A., & Ghazisaedy, M. (2015). The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition. *International Journal of Social Robotics*, *7*(4), 523–535. https://doi.org/10.1007/s12369-015-0286-y

Ayedoun, E., Hayashi, Y., & Seta, K. (2019). Adding communicative and affective strategies to an embodied conversational agent to enhance second language learners' willingness to communicate. *International Journal of Artificial Intelligence in Education*, *29*(1), 29–57. https://doi.org/10.1007/s40593-018-0171-6

Belpaeme, T., Vogt, P., van den Berghe, R., Bergmann, K., Göksun, T., de Haas, M., Kanero, J., Kennedy, J., Küntay, A. C., Oudgenoeg-Paz, O., Papadopoulos, F., Schodde, T., Verhagen, J., Wallbridge, C. D., Willemsen, B., de Wit, J., Geçkin, V., Hoffmann, L., Kopp, S., … Pandey, A. K. (2018). Guidelines for designing social robots as second language tutors. *International Journal of Social Robotics*, *10*(3), 325–341. https://doi.org/10.1007/s12369-018-0467-6

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

Bibauw, S., François, T., & Desmet, P. (2015). Dialogue-based CALL: an overview of existing research. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds.), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 57–64). Research-publishing.net. https://doi.org/10.14705/rpnet.2015.000310

Bibauw, S., Van den Noortgate, W., François, T., & Desmet, P. (2022). Dialogue systems for language learning: A meta-analysis. *Language Learning & Technology*, *26*(1), to be published.

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

Cornillie, F., Lagatie, R., Vandewaetere, M., Clarebout, G., & Desmet, P. (2013). Tools that detectives use: In search of learner-related determinants for usage of optional feedback in a written murder mystery. In P. Hubbard, M. Schulze, & B. Smith (Eds.), *Learner-Computer Interaction in Language Education: A Festschrift in Honor of Robert Fischer* (pp. 22–45). CALICO.

Dizon, G. (2020). Evaluating intelligent personal assistants for L2 listening and speaking development. *Language Learning & Technology*, *24*(1), 16–26. https://doi.org/10125/44705

Engwall, O. (2012). Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. *Computer Assisted Language Learning*, *25*(1), 37–64. https://doi.org/10.1080/09588221.2011.582845

Engwall, O., & Lopes, J. (2020). Interaction and collaboration in robot-assisted language learning for adults. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2020.1799821

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, *51*(10), 832–844. https://doi.org/10.1016/j.specom.2009.04.005

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. https://doi.org/10125/44719

Gass, S. M., & Mackey, A. (2015). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 180–206). Routledge.

González-Lloret, M., & Ortega, L. (2014). Towards technology-mediated TBLT: an introduction. In M. González-Lloret & L. Ortega (Eds.), *Technology-mediated TBLT: Researching technology and tasks* (pp. 1–22). John Benjamins.

Han, J. (2012). Robot assisted language learning. *Language and Learning Technology*, *16*(3), 1–9.

Heift, T., & Schulze, M. (2015). Research timeline: Tutorial computer-assisted language learning. *Language Teaching*, *48*(4), 471–490. https://doi.org/10.1017/S0261444815000245

Holland, V. M., Kaplan, J. D., & Sams, M. R. (Eds.). (1995). *Intelligent language tutors: Theory shaping technology*. Lawrence Erlbaum.

Hong, Z.-W., Huang, Y.-M., Hsu, M., & Shen, W.-W. (2020). Authoring robot-assisted instructional materials for improving learning performance and motivation in EFL classrooms. *Educational Technology & Society, 19*(1), 337–349.

Hsu, L. (2020). To CALL or not to CALL: Empirical evidence from neuroscience. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2020.1750429

Jokinen, K., & McTear, M. F. (2010). *Spoken dialogue systems*. Morgan & Claypool. https://doi.org/10.2200/s00204ed1v01y200910hlt005

Jurafsky, D., & Martin, J. H. (2020). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). Self-published draft. https://web.stanford.edu/~jurafsky/slp3/

Lee, S., Noh, H., Lee, J., Lee, K., Lee, G. G., Sagong, S., & Kim, M. (2011). On the effectiveness of robot-assisted language learning. *ReCALL, 23*(1), 25–58. https://doi.org/10.1017/s0958344010000273

Litman, D., Strik, H., & Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*, *15*(3), 294–309. https://doi.org/10.1080/15434303.2018.1472265

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–452). Oxford University Press.

McMillan, S. J. (2002). A four-part model of cyber-interactivity: Some cyber-places are more interactive than others. *New Media & Society, 4*(2), 271–291. https://doi.org/10.1177/14614440222226370

Robinson, P., & Gilabert, R. (2007). Task complexity, the Cognition Hypothesis and second language learning and performance. *International Review of Applied Linguistics in Language Teaching, 45*(3), 161–176. https://doi.org/10.1515/iral.2007.007

Sato, M., & Lyster, R. (2012). Peer interaction and corrective feedback for accuracy and fluency development: Monitoring, practice, and proceduralization. *Studies in Second Language Acquisition*, *34*(4), 591–626. https://doi.org/10.1017/S0272263112000356

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., & Pineau, J. (2018). A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, *9*(1), 1–49.

Sydorenko, T., Daurio, P., & Thorne, S. L. (2018). Refining pragmatically-appropriate oral communication via computer-simulated conversations. *Computer Assisted Language Learning*, *31*(1–2), 157–180. https://doi.org/10.1080/09588221.2017.1394326

Sydorenko, T., Smits, T. F. H., Evanini, K., & Ramanarayanan, V. (2019). Simulated speaking environments for language learning: Insights from three cases. *Computer Assisted Language Learning*, *32*(1–2), 17–48. https://doi.org/10.1080/09588221.2018.1466811

Taguchi, N., Li, Q., & Tang, X. (2017). Learning Chinese formulaic expressions in a scenario-based interactive environment. *Foreign Language Annals, 50*(4), 641–660. https://doi.org/10.1111/flan.12292

Thomas, M., Reinders, H., & Warschauer, M. (2012). Contemporary computer-assisted language learning: The role of digital media and incremental change. In M. Thomas, H. Reinders, & M. Warschauer (Eds.), *Contemporary computer-assisted language learning* (pp. 1–12). Bloomsbury.

Timpe-Laughlin, V., Sydorenko, T., & Daurio, P. (2020). Using spoken dialogue technology for L2 speaking practice: What do teachers think? *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2020.1774904

van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., van der Ven, S., & Leseman, P. (2019). Social robots for language learning: A review. *Review of Educational Research*, *89*(2), 259–295. https://doi.org/10.3102/0034654318821286

Wilske, S. (2015). *Form and meaning in dialog-based computer-assisted language learning* [Doctoral dissertation, Universität des Saarlandes]. https://doi.org/10.22028/D291-23654

Ziegler, N. (2016). Taking technology to task: Technology-mediated TBLT, performance, and production. *Annual Review of Applied Linguistics*, *36*, 136–163. https://doi.org/10.1017/S0267190516000039