Discussing with a computer to practice a foreign language: research synthesis and conceptual framework of dialogue-based CALL

Serge Bibauw[1,2,3,4], Thomas François[3], and Piet Desmet[4]

[1] ITEC, KU Leuven, Kortrijk, Belgium

[2] IMEC, Leuven, Belgium

[3] CENTAL, UCLouvain, Louvain-la-Neuve, Belgium

[4] Universidad Central del Ecuador, Quito, Ecuador

Abstract

This chapter presents the results of a systematic review of the literature on dialogue-based computer-assisted language learning (CALL), resulting in a conceptual framework for research on the matter. Applications allowing a learner to have a conversation in a foreign language *with* a computer have been studied from various perspectives and under different names (dialogue systems, conversational agents, chatbots...). Considering the fragmentation of what we identify under the term dialogue-based CALL, we attempt to offer a structured overview of these efforts into a conceptual framework. Through a methodical search strategy, we collected a corpus of 343 publications. From this corpus, we formalized an operational definition of dialogue-based CALL, which allowed us to identify 96 relevant systems. Analysing the type of dialogue they offer, on a continuum of constraints on form and meaning, we propose to classify those systems into four groups. We have called these *branching*, *form-focused*, *goal-oriented* and *reactive* systems, and we describe their corresponding interactional, instructional and technological traits. We summarise the main results from empirical studies on such systems, distinguishing observational, survey and experimental studies, and discuss the impact of dialogue-based CALL on motivation and L2 development, identifying positive evidence on both outcomes. Finally, we propose two main avenues for future research: relative effectiveness of dialogue-based CALL approaches, and dialogue systems as an environment for testing second language acquisition (SLA) hypotheses.

This is the Accepted Manuscript of an article published in *Computer Assisted Language Learning* on 14/02/2019, available online: http://www.tandfonline.com/10.1080/09588221.2018.1535508.

To cite this article:
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*, [*FirstView*], doi:10.1080/09588221.2018.1535508.

CONTENTS

1 Introduction 3
1.1 A dispersed and fragmented field . . . 3
1.2 Rationale for dialogue-based CALL . . . 4
1.3 Research questions . . . 5
2 Methodology 5
2.1 Search strategy . . . 5
2.2 Inclusion and exclusion criteria . . . 6
3 Delineating the field 7
3.1 An operational definition . . . 7
3.2 Applying the definition to our dataset . . . 8
3.3 An emerging research domain . . . 9
3.3.1 ICALL: focus on written output correction . . . 9
3.3.2 CAPT: focus on pronunciation scoring and correction . . . 10
3.3.3 SDS and conversational agents: focus on dialogue management . . . 10
3.3.4 Chatbots: focus on reactive response selection . . . 11
3.3.5 Convergence and recent tendencies . . . 11
3.4 Representing the field's evolutions and tendencies . . . 12
4 Towards a typology of dialogue-based CALL systems 15
4.1 Constraints on learner production . . . 16
4.2 A typology of dialogues and systems . . . 16
4.3 Instructional characteristics . . . 19
4.4 Interactional implications . . . 20
4.4.1 Initiative management . . . 20
4.4.2 Goal-orientation . . . 21
4.5 Technological implications . . . 21
4.5.1 Variation, predictability and processing . . . 21
4.5.2 Dialogue management and natural language understanding . . . 22
4.6 Summary: types and constraints as design choices . . . 23
5 A synthesis of effectiveness studies' results 24
5.1 Types of studies . . . 24
5.2 Observational studies . . . 26
5.3 Survey studies . . . 27
5.4 Effectiveness studies . . . 28
5.4.1 Effects on motivation . . . 28
5.4.2 Effects on language development . . . 28
6 Conclusions and avenues for research 30

Acknowledgements 32

References 32

Supplementary Annexes 43

1 INTRODUCTION

Since the beginning of the 1980s, researchers and developers have attempted to develop systems allowing learners to practice a second or foreign language (L2) through meaningful conversational interactions with a computer, in order to develop their L2 proficiency. While computer-mediated communication (CMC) explores the way language users can interact *through* a computer, here, we focus on autonomous systems where the computer *is* the interlocutor. Such efforts have been made from different backgrounds and perspectives, and under many different names, from *chatbots* and *conversational agents* to *robots* and *dialogue systems*. This chapter attempts to present a consolidated overview of these efforts under the umbrella term *dialogue-based CALL*, and to develop a conceptual framework for research on the topic. We also intend to set a research agenda by examining what types of studies and empirical designs have been used in this field and what insights have already been gained on the effectiveness of dialogue-based CALL.

1.1 A dispersed and fragmented field

Studies on dialogue-based CALL have traditionally been scattered among different categories, as it appears in the literature reviews of CALL mentioning such learning environments (Eskenazi, 2009; Golonka, Bowles, Frank, Richardson & Freynik, 2014; Wachowicz & Scott, 1999): these systems appear under *intelligent tutoring systems* (*ITS*) when they offer customised written instruction, *automatic speech recognition* (*ASR*)*-based CALL* or *computer-assisted pronunciation training* (*CAPT*) when they involve oral interaction or pronunciation training, or under *virtual worlds* and *serious games* when the dialogues are contextualised in a broader narrative. At a global level, dialogue-based CALL has often been divided into *spoken* systems, mostly struggling to improve speech recognition, and *written* systems, mainly concerned with error diagnosis, as if these were the only two natural language processing (NLP) problems at stake.

This situation has had two repercussions. First, research on dialogue-based CALL has been dispersed, limited to small clusters of projects, with researchers often unaware of the existence of similar efforts happening in other traditions (Bibauw, François & Desmet, 2015). Second, NLP challenges related to dialogue management on the semantic (natural language understanding, natural language generation) and pragmatic (dialogue act recognition, dialogue modelling, grounding...) levels, although crucial for language learning, have been systematically overlooked in the CALL literature, while the NLP literature disregarded the importance of the instructional and interactional design of such interactions.

Our claim is that, across the various traditions and terms, beyond the multiple forms the interaction might take and the numerous technologies to tackle it, dialogue-based CALL corresponds to a consistent undertaking, i.e. allowing a learner to practice an L2 autonomously in meaningful conversations. These systems face many similar technological and instructional challenges, and would benefit from combined efforts in research. They also share a common rationale.

1.2 Rationale for dialogue-based CALL

Dialogue-based CALL efforts share the broad assumption that meaningful practice of a target language, as it occurs in conversation, leads to improve the learner's proficiency in that language, and that, even if a native speaker remains the ideal interlocutor, a computer can provide opportunities for such practice (e.g. Seneff, Wang, Peabody & Zue, 2004).

Dialogue-based CALL finds a prominent foundation in the interactionist perspective on SLA (Long, 1996), as dialogue naturally offers opportunities for input, output and interaction. The automated agent provides *input*, whose complexity can be adjusted to the learner level. On every other turn, the learner has to express their intended meaning, which can be seen as an instance of *pushed output* (Swain, 2005). Moreover, the fact that the written transcription of the dialogue is often visible for the learner promotes *noticing*, both of their own errors and of new structures present in the input (Lai & Zhao, 2006). The major value of dialogue lies in the *interaction* it offers with the other speaker, and especially in instances of *negotiation of meaning* and *feedback*, which help learners notice the gap between their production and the target structures (Pica, 2013). There is now an important body of research supporting the fact that interaction itself conveys actual learning (Mackey & Goo, 2007; Plonsky & Gass, 2011), and that computer-mediated interaction provides the majority of the benefits ascribed to the interaction hypothesis (e.g. Jepson, 2005). Lastly, dialogue-based CALL allows for the *proceduralisation* of existing knowledge, by automatizing linguistic routines (DeKeyser, 2007), and thus to 'develop learners' spontaneous productive skills' and L2 fluency (Muranoi, 2007, p. 55).

In many foreign language learning contexts, students lack occasions to use the L2 outside of the classroom (Fryer & Carpenter, 2006), and even inside of it, spoken interactive practice is often confined to teacher-learner interactions, and limited by large class sizes or by first language (L1) use (Ortega, 2007). In such contexts, dialogue-based CALL provides an opportunity to practice meaningful conversations in a kind of 'virtual immersion' that, although not necessarily as effective as an interaction with a native speaker, may offer many of its characteristics (N. C. Ellis & Bogart, 2007). Dialogue-based CALL can also offer opportunities for spontaneous interactive L2 production for participants of MOOCs, and online language learning in general, that often lack such activities, in particular for oral skills (Read, 2014).

And it may even provide some advantages over human interlocutors. First, dialogue-based CALL systems are available at any moment for as long as the learner wishes to practice. They do not object to repeating the same interaction and do not lose their patience in front of a struggling speaker (Fryer & Carpenter, 2006). Because the learners are conscious of the artificiality of the agent, such systems offer a low-anxiety environment for practice, which can positively affect learners' willingness to communicate (Ayedoun, Hayashi & Seta, 2015).

Finally, they offer a fully controllable learning environment, potentially configurable towards optimal conditions on all impacting factors (feedback, learner modelling and adaptivity, motivational support, etc.), for learning, but also for research purposes. By avoiding the unpredictable variation of a human interlocutor, dialogue-based CALL can indeed offer fully monitored conditions for conducting empirical research on L2 interaction (Hegelheimer & Chapelle, 2000).

1.3 Research questions

In this chapter, we propose a common framework for research on dialogue-based CALL. Through a systematic research synthesis, we attempt to answer three research questions:

RQ1.1 What are the boundaries of the field of dialogue-based CALL, how can we define this field, and what have been its major traditions and evolutions?

RQ1.2 How can we categorize and distinguish the different types of dialogue-based CALL systems that have been developed so far, from interactional, instructional and technological perspectives?

RQ1.3 What types of research and empirical designs have been used to study the impact of dialogue-based CALL, and what insights have been gained on its effectiveness?

To answer these questions, we conducted a systematic research review, intending to gather all relevant research on the subject, whose methodology is detailed below. From this data, we attempt to define the scope of dialogue-based CALL and formulate an operational definition, allowing us to determine more precisely the different research and technological subfields, as well as the chronological evolutions and current tendencies in the domain. Furthermore, we draw a general bottom-up typology of dialogue-based CALL systems, structured on interactional, instructional and technological criteria. Finally, we summarise the various empirical effectiveness studies conducted on dialogue-based CALL and identify the research challenges that remain to be addressed.

2 METHODOLOGY

In order to obtain a better understanding of previous research and developments in dialogue-based CALL, we conducted a systematic review of the existing literature.

2.1 Search strategy

First, we carried out a replicable, exhaustive search on three meta databases: Thomson Reuters' Web of Science (databases included: Web of Science Core Collection and INSPEC), ProQuest (databases included: ABI/INFORM, ERIC, International Bibliography of the Social Sciences, Linguistics and Language Behaviour Abstracts and Periodicals Archive Online) and Elsevier's Scopus. The search syntax combined all the terms identified as potential keywords for dialogue-based CALL (see Bibauw et al., 2015, for a discussion of these keywords) with a set of common terms referring to language learning:

```
(chatbot / chat bot / chatterbot / conversational agent /
conversational companion / conversational system /
dialog* system / dialog* agent / dialog* game /
pedagogical agent / human-computer dialog* / dialog*-based)
+ ((language / English) (learning / teaching / acquisition)
/ (second / foreign) language / L2 / EFL / ESL / ICALL)
```

It was looked up on titles, abstracts and keywords. It gathered respectively 99 hits on Web of Science, 129 on Scopus and 12 on ProQuest, with some overlap between them, thus resulting in the identification of 159 papers.

As a secondary search strategy, from all the relevant references found in the primary recollection, we reviewed forward citations (new publications citing reference) and realized an ancestry search (older publications cited by reference). This step was particularly important considering the important disparity in terms and concepts' use across the various fields where dialogue-based CALL appear, and the fact that many relevant publications are absent from the above-mentioned databases. It added 184 more papers to our collection, totalling 343 documents[1].

2.2 Inclusion and exclusion criteria

Previously found documents were systematically reviewed and coded regarding the characteristics of the research and the system(s) presented. Only documents satisfying the following eligibility criteria were included:

(1) The presented system or application involved interactions in natural language with some form of computer or automated agent (this voluntarily broad definition will be refined in the next section).

(2) Second language learning was the design goal of the system or of the study. A certain number of publications identified by the search were thus excluded from our study because language learning was only mentioned as one of the potential fields of application (e.g. Griol, Baena, Molina & Sanchis de Miguel, 2014). We also left apart a few studies applied to primary language acquisition, either for children (e.g. Y. Kim, 2013) or to adult communicative skills development (Vaassen et al., 2012).

(3) The above-mentioned system or its application to language learning was the main focus of the publication. This excluded certain papers that only mentioned the existence or the possibility of a dialogue system (e.g. Lorenzo, Lezcano & Sánchez-Alonso, 2013), presented a technological component, such as a parser or a dialogue manager, but whose application to dialogue-based CALL was not discussed (e.g. Chen & Tokuda, 2003), as well as reviews of CALL that only briefly mentioned dialogue applications.

(4) The document was a peer-reviewed publication — papers published in a peer-reviewed journal or presented at an international conference, or chapter in an edited book —, or a doctoral dissertation.

Besides, we also had to exclude at this stage a few papers that could not be accessed online or in major university libraries, papers written in languages we could not understand (Korean, Chinese), and a couple of duplicate versions of papers that were already included (republications).

After the inclusion and exclusion process, we obtained a final pool of 250 publications, ranging from 1982 to June 2017.

[1] The complete pool of publications is provided in Annex I.

3 DELINEATING THE FIELD

3.1 An operational definition

From the systematic review of our corpus of studies, we propose an operational definition of dialogue-based CALL as any system or application where the activity consists for the learner to engage in a dialogue with an automated interlocutor in a L2.

Firstly, dialogue-based CALL is thus characterised by the fact that the interacting agent, i.e. the communicational counterpart of the learner, is a virtual agent controlled by the computer. The learn interacts *with* the computer. This excludes the conversational activities carried out with another human *via* a computer, usually referred to as *computer-mediated communication*, which have been abundantly studied in CALL since the 1990s (see Ziegler, 2016). As the system plays the role of interlocutor, and sometimes also tutor, and as the learner practices individually, dialogue-based CALL is clearly a form of *tutorial CALL* (Heift & Schulze, 2015). A few systems supplement interactions between learners (CMC) with the tutoring of an automated agent: in TutorBot (Lu, Chiou, Day, Ong & Hsu, 2006), MentorChat (Tegos, Demetriadis & Karakostas, 2013) and PASCALL (da Costa Pinho, Epstein, Reategui, Corrêa & Polonia, 2013), for instance, discussions between L2 learners are guided by prompts and feedback from a pedagogical agent. However, as they cannot qualify as a strict learner-computer interaction, we decided to leave such computer-supported collaborative learning (CSCL) systems out of the scope of dialogue-based CALL.

Secondly, whereas most practice in tutorial CALL is *item-based*, dialogue-based CALL is organized around the *dialogue* as unit of instruction. This fundamentally differentiates it from production activities revolving around isolated items, most often equivalent to a sentence, as found in most language courseware (Heift & Schulze, 2015). Rather than being a syntactic unit, a dialogue is a pragmatic unit, involving interactional strategies and various complex phenomena that language learners must acquire to develop communicative competence (Kormos, 1999). In dialogue-based CALL, the meaning is co-constructed through various conversational turns. Consequently, systems where an agent is used to help practice isolated and self-contained fragments (e.g. Griol et al., 2014; Massaro, Liu, Chen & Perfetti, 2006), without a sequence of turns, cannot be considered to be based on dialogue.

Finally, in dialogue-based CALL the dialogic interaction in the L2 constitutes the task itself. This excludes systems that only deliver instruction or commands through an agent, without asking for any verbal response from the user, commonly referred to as *pedagogical agents* (e.g. Bergmann & Macedonia, 2013; Gupta, Walker & Romano, 2008), or tutorial dialogue in the native language of the learner (e.g. Saerbeck, Schut, Bartneck & Janse, 2010).

We synthesize how dialogue-based CALL is distinct from dialogue systems in general, computer-mediated communication CALL, item-based tutorial CALL, and pedagogical agents in Table 1.

This definition is deliberately inclusive, as it matches all chatbots, dialogue systems and conversational agents used for language learning, but also applies to certain applications that, while they make use of dialogues, do not necessarily involve complex NLP to analyse the learner output or to generate the agent response. Such systems either do not adapt the content of the dialogue to the user's actions, by following a predetermined script (e.g. Cornillie, Lagatie, Vandewaetere, Clarebout & Desmet, 2013; Kwon et al., 2015; Levin & Evans, 1995), or limit the user output to a closed set of possibilities, by providing a

	CMC CALL	ITEM-BASED TUTORIAL CALL	PEDAGOGICAL AGENTS IN CALL	DIALOGUE-BASED CALL
Interlocutor	Human	System	System	System
Interactional unit	Dialogue	Item	Dialogue	Dialogue
Role of interaction	Task	Task	Scaffolding	Task

Table 1 – Dialogue-based CALL's defining criteria as they distinguish it from other CALL subdomains.

list of words to be used (e.g. Krüger & Hamilton, 1997) or a list of utterances to be chosen from — a system known as *branching dialogue*, which is frequently used in adventure games, and sometimes in CALL (e.g. Stewart & File, 2007). Yet, because these restrictions can be seen as strategies designed to cope with the challenges of managing automated conversations, we consider that such systems do account for a certain type of dialogue-based CALL.

3.2 Applying the definition to our dataset

We used our operational definition to analyse each publication from our corpus. After filtering out the publications and systems that could not be considered as dialogue-based CALL — either for involving human interlocutors, item-based activities or dialogue only as a scaffolding strategy —, 207 publications remained from the original 250[2].

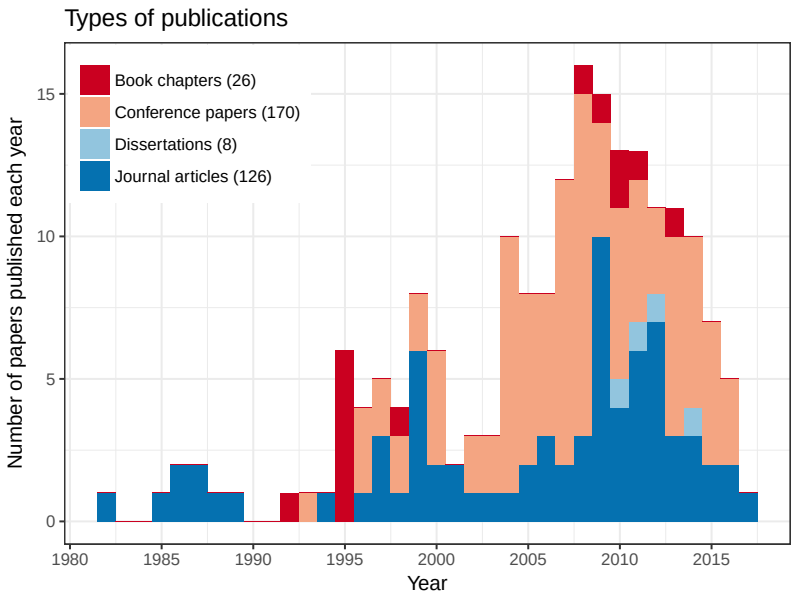

Figure 1 – Published papers on dialogue-based CALL grew in numbers, especially in the last decade.

Figure 1 presents a chronological evolution of publications in our dataset that fit the definition of dialogue-based CALL. As it clearly shows, most of these publications are journal articles (mainly in CALL, SLA or educational technology journals) and conference papers (mainly at artificial intelligence

[2] The complete list of those papers, with their coding variables, is presented in Annex II.

and NLP conferences). This fact reveals that the topic of dialogue-based CALL has been explored in both applied linguistics and NLP. It is also clear that the topic is still relatively young, being on the upswing since 2007.

3.3 An emerging research domain

Our corpus of studies shows that dialogue-based CALL has appeared through the years in different fields and research traditions, based on different technologies, and aiming at different objectives. Four major strands can be identified: (1) 'intelligent' CALL (ICALL), (2) computer-assisted pronunciation training (CAPT), (3) spoken dialogue systems and conversational agents (SDS/CA), and (4) chatbots. Table 2 illustrates each of these strands with a dialogue excerpt from one of the systems.

SYSTEM	TE KAITITO	CANDLETALK	SPELL	JABBERWACKY
REFERENCE	Vlugter et al., 2009	Chiu et al., 2007	Morton & Jack, 2005	Fryer & Nakao, 2009
STRAND	ICALL	CAPT	SDS/CA	Chatbot
MODALITY	Written	Spoken	Spoken	Written
CONTEXT	User is discussing with three agents, to practice personal pronouns. *Translated from Maori.*	User is loudly playing videogames. System is his roommate, who comes to talk to him.	User is with another character in a café. They are ordering food.	No context is given for the conversation. User is free to bring up any topic. System responds.
DIALOGUE EXCERPT	**S1**: Where are you from? **U**: I'm from Dunedin. [error] **S1**: There's a mistake in that sentence. Maybe you mean 'I'm from Dunedin'. Let's try again. Where are you from? **U**: I'm from Dunedin. [no error] **S1**: Well done! **S2**: I'm from Dunedin too. **S3**: Let's see if you remember... Where are you and S2 from? **U**: You two are from Dunedin.	**U**: [*choosing from a list of utterances*] Wow! What a great game! **S**: Hey! Your computer is too loud. Can you please turn it down? **U**: [*choose one of the following sentences and pronounce it*] - What? What sound? I didn't hear anything. - Pardon me; what did you say? - Oh, I'm sorry. I was concentrating on the game so I didn't notice. Did I bother you? (...)	**S**: What food does Katie like? **U**: [silent] **S**: What food does Katie like? [slower] **U**: Umm—food... **S**: John likes sushi. What food does Katie like? **U**: Umm Katie like pizza. **S**: That's right. Katie likes pizza. What food do you like?	**U**: Let's talk about character! **S**: No, let's talk about politics. **U**: I think that I have a character hating to lose. **S**: Everyone has a bad period from time to time. **U**: And I am very careless. Are you? **S**: No. **U**: Good. There are many careless people in my friend. **S**: There are many problems too.

Table 2 – Exemplars of dialogue from four systems, illustrating each strand of dialogue-based CALL. Dialogue excerpts are adapted from the original publications.

3.3.1 ICALL: focus on written output correction

In the early 1980s, insights gained from previous CALL development efforts (Hart, 1981), together with the influence of Krashen's (1982) theories and the rise of the communicative approach to language teaching, encouraged researchers to set new goals for CALL. Underwood (1984) was the first to insist on developing 'communicative CALL', an approach that would take advantage of novel artificial intelligence techniques to implement 'meaningful' communication practice in tutorial systems, through conversational interaction, among other activities. Underwood (1982) developed a written conversational program for Spanish, FAMILIA, whose pattern matching functioning was largely inspired by the first chatbot, ELIZA (Weizenbaum, 1966).

This effort towards more meaning-focused activities coincided with a plead for more 'intelligent' automatic processing of learner responses, under the flagship of ICALL. It is especially the implementation of automatic corrective feedback (NLP) within a meaning-focused communicative activity (commu-

nicative CALL) that interested many CALL developers. The dialogue with an agent was seen as a preferential setting for such endeavour, and implemented in various intelligent tutoring systems: FAMILIA, L2TUTOR (Price, Bunt & McCalla, 1999), SUBARASHII (Bernstein, Najmi & Ehsani, 1999).

However, the open-endedness of conversations multiplied exponentially the number of possible user entries the system had to process. For this reason, various researchers integrated the constraints from text adventure games in CALL games, such as SPION (Molla, Sanders & Sanders, 1988) and FLAP (Culley, Mulford & Milbury-Steen, 1986), and *microworlds*, such as FLUENT (Hamburger & Hashim, 1992), LINGO (Felshin, 1995) and the MILITARY LANGUAGE TUTOR (Holland, Kaplan & Sabol, 1999). Contrary to a conversation, here the user gives orders in natural language ('Open the door'), and the system responds ('You don't have the key'). Another option to reduce the unpredictability of user entries is to have the system controlling the conversational flow (*system-initiated*, see 4.4.1), e.g. by asking questions to the user, which works well if the objective is essentially to provide corrective feedback, as in systems such as TE KAITITO (Vlugter, Knott, McDonald & Hall, 2009) and GENIETUTOR (Kwon et al., 2015).

All those CALL systems were mainly referred to as 'intelligent language tutors' (Holland, Kaplan & Sams, 1995), and constituted the initial essence of ICALL. However, considering the state of the art in NLP at that time, most research efforts in ICALL addressed the NLP issues of parsing language learners productions, especially for error diagnosis and feedback (e.g. DeSmedt, 1995; Feuerman, Marshall, Newman & Rypa, 1987). Very few addressed the actual dialogue management issues, the instructional design of conversational tasks or their effectiveness for learning.

3.3.2 CAPT: focus on pronunciation scoring and correction

At the end of the 1990s, when speech processing started showing promising results, researchers began to look into ways to analyse and provide feedback on *spoken* output. CAPT programs are, as their written counterparts, mostly item-based, but various researchers tried to integrate this spoken practice into dialogues. Most developments in spoken dialogue-based CALL originated from team specialized in speech technologies, e.g. VILTS/ECHOS (Rypa, 1996; Rypa & Price, 1999), FLUENCY (Eskenazi & Hansma, 1998), ARTUR (Engwall, 2012) and GREET (Cucchiarini, Bodnar, Penning de Vries, van Hout & Strik, 2014). The use of dialogue by most of these systems is only partial, and the liberty of the user to navigate the dialogue is restricted, as the focus remains on evaluation and feedback provision. However, these efforts paved the way of spoken interaction for spoken dialogue system (SDS), which would a few years later revisit dialogue management.

3.3.3 SDS and conversational agents: focus on dialogue management and multimodality

Spoken dialogue systems appeared in the 1970s as telephone interfaces for customer services, and have been vastly researched since, in the field of speech technology. The domain has been the source of important advancements in our understanding of dialogue, which have led to more complex dialogue management processes, involving dialogue state monitoring, initiative management, and natural language understanding (Jokinen & McTear, 2010). The availability of SDS frameworks, such as GALAXY (Seneff et al., 1998), allowed to develop new applications for language learning (Raux & Eskenazi, 2004; Seneff, Wang & Chao, 2007).

In parallel, the generalisation of web interfaces added visual and text support for previously audio-only services. Hence, the spoken mode became less distinctive, as ASR and text-to-speech (TTS) modules allowed to quickly pass from one mode to the other, and interfaces became more multimodal, involving often an avatar, or (*embodied*) *conversational agent*, capable of gestures, facial and body expressions (Cassell, Bickmore, Campbell, Vilhjálmsson & Yan, 2000).

Some advanced dialogue-based CALL applications have been developed in these areas, offering task-oriented interactions with embodied agents, contextualised in 3D virtual worlds, such as SPELL (Anderson, Davidson, Morton & Jack, 2008), DEAL (Hjalmarsson, Wik & Brusk, 2007) and Alelo's TLCTS (Johnson & Valente, 2009). These systems differ fundamentally from ICALL and CAPT ones, as they put the emphasis on the construction of the conversation itself, rather than isolating target structures and providing feedback. The user has more control over the semantic content of its intervention, which is a precondition to *focus on meaning*.

3.3.4 Chatbots: focus on reactive response selection

Meanwhile, in a different context, the pioneering ELIZA also constituted the starting point of many efforts on chatbots, text-based dialogue systems. These developments were accelerated at the end of the 1990s by the release of Artificial Intelligence Markup Language (AIML), a mark-up language and framework making the programming of chatbots more accessible (Wallace, 2003). The availability of an open source framework and a vast community of developers allowed for some CALL practitioners either to use existing (general purpose) chatbots, analysing their potential for language learning (e.g. Coniam, 2008, 2014; Fryer & Nakao, 2009), or to create new systems intended for language learners, such as CSIEC (Jia, 2009) or TutorBot (Lu et al., 2006).

Still, despite their initial appeal, the majority of these text systems, functioning on handcrafted pattern-matching rules, present strong limitations, both technical and pedagogical (see Sha, 2009; Williams & van Compernolle, 2009). Among other shortcomings, too much attention may have been given to the challenge of responding to almost anything, and not enough to dialogue management, leading virtually all chatbots to be only reactive and to completely omit goal-oriented dialogue. The research in that particular area is also relatively disconnected from the literature on ICALL (see 3.4), and from the technological developments in NLP.

3.3.5 Convergence and recent tendencies

In the last ten years, technological advancements from SDS and pedagogical awareness from ICALL seem to have converged into new systems, exploiting complex NLP approaches, combined with more precise learning goals and sounder evaluation methods. Systems such as Sasha (Petersen, 2010), while still focusing on corrective feedback, present complex dialogue management procedures, and very thorough effectiveness evaluations (see also Wilske, 2015). It is also the case for SDS implemented in virtual worlds, with a gaming component, as POMY (K. Lee, Kweon, Lee, Noh & Lee, 2014) and IVELL (Hassani, Nahvi & Ahmadi, 2016), and some SDS implemented into physical robots (S. Lee et al., 2011). These systems already display, in parts, features that will be of major importance for future developments in dialogue-based CALL: multimodality, meaningful and authentic contexts of communication, goal-oriented interactions, mixed initiative, and complex dialogue management.

3.4 Representing the field's evolutions and tendencies

To provide an empirically founded representation of the field, we performed a co-citation network analysis on our set of publications. Co-citation analysis establishes a matrix based on how frequently two authors are cited together, which is then used for social network analysis (Otte & Rousseau, 2002). The objective of the analyses is to provide a data-driven overview of the communities of research within a field, as well as identifying authorities and general tendencies.

For practical reasons, and because of the narrowness of the field, we counted frequency of co-citation — i.e. when two items are present together in the same document — for first author last names[3] and for system names. The frequency counts were computed automatically, by pattern matching, on the full-text version of the papers from our corpus. The inevitable false positives (mainly namesakes) were corrected manually. False negatives (missed occurrences) are relatively rare, only due to optical character recognition (OCR) errors in a minority of documents. It should be emphasised that, to allow the automatic counting of occurrences, we did not compute the co-citation of *references* but of *first authors* (e.g. references to all papers with S. Seneff as first author are taken into account under the author key `Seneff`).

Remain only two limitations. 1. Some last names are shared by various authors of our corpus (e.g. `Lee`, `Price`, `Wang`); in most cases, only one was a first author, and this is the one we considered in our manual disambiguation; in cases with various first authors with the same name, we selected the most represented one and ignored the subsequent namesakes, to avoid amalgamating their connections. 2. Regarding systems, some of the authors have not given a specific name to their system (e.g. Ayedoun et al., 2015; Wilske, 2015) and can thus not be referred to by their name; those programs are thus omitted in the network of systems.

The co-citation counts were then processed to produce a network representation (Figures 2 and 3). Each node, representing either an author or a system, is connected to another if both are cited in the same paper; the strength of the connection (i.e. the number of times two items are cited together) is illustrated by the weight of the line, and the relative importance of each node (i.e. how frequently is it co-cited) by its size.

Besides providing a data-driven representation of the relationships and the relative influence of each author and system, the co-citation graphs make clearly apparent certain tendencies and trends of the domain. In the network of authors, a diagonal drawn from Underwood to Lee corresponds more or less to a continuum written-spoken, with multimodal systems in the crowded centre Pools of researchers working on similar issues are close-by.

On the other hand, the authors and systems from the chatbots strand are clearly less connected, both together and with the rest of the literature: only a few nodes (`Coniam`, `Jia`) make the link between chatbots and the other systems. Several systems and authors are even completely isolated (not represented on the graph, see footnote), because their work is never cited along with others and they do not refer to other dialogue-based CALL authors in their own articles.

The graph also confirms the existence of different groups of research: ICALL researchers and systems are grouped together, with SDS/CA on the other side.

[3] For the co-citation of authors, we also considered the citing author, as it clearly established that he was inspired by the cited author. Hence, e.g. when Wilske (2015) cites both Vlugter (from Vlugter et al., 2009) and Morton (from Morton, Davidson and Jack, 2008), three connections are computed: `Vlugter--Morton` (the co-citation per se), but also `Wilske--Vlugter` and `Wilske--Morton`.

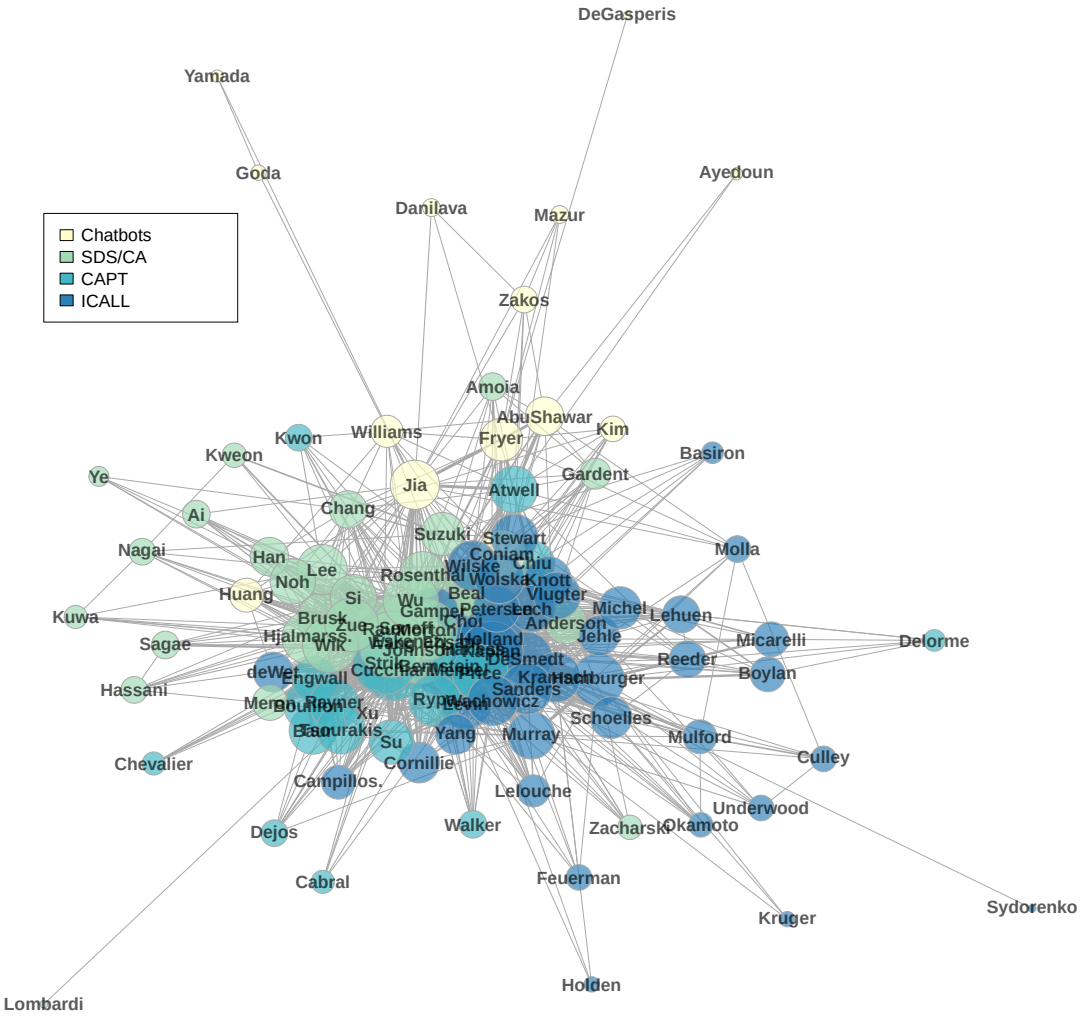

Figure 2 – Co-citation network of first authors present a strong connection between the ICALL, CAPT and SDS/CA traditions. Nodes placement on the graphs calculated by the force-directed layout algorithm (Fruchterman & Reingold, 1991). Isolated nodes (mentioned only in their own publication(s), without any other author from our corpus mentioned) are not drawn. List of isolated author nodes: `Chatterjee`, `Cho`, `Harroff`, `Kondo`, `Powers`, `Sha`, `Tanghe`.

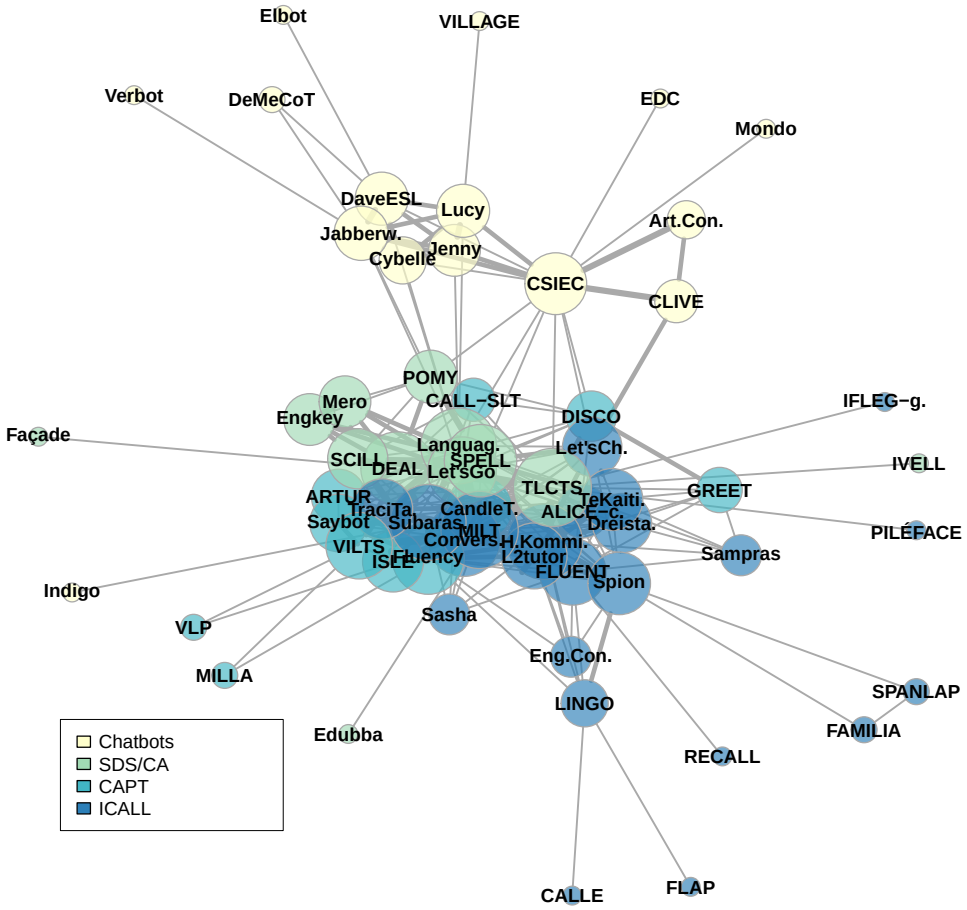

Figure 3 – Co-citation network of systems shows a stronger separation between the main cluster of SDS/CA and ICALL systems, and the chatbots strand. Isolated nodes (mentioned only in their own publication(s)) are not drawn. List of isolated system nodes: Aghate, Chat Bots Mediator, Conversation Rebuilding, Die Sprachmaschine, Fable expert, GenieTutor, JDT, Mentira, PETA, Request game, SimCon and Sprinter.

Those two traditions that have been brought closer and linked across the years by the convergence of technologies and interests. A chronological splitting of the network shows how ICALL was dominant until 2000, when CAPT efforts appeared, with then a clear intensification of research in SDS/CA. The fact that, in 2017, the network involving ICALL, CAPT and SDS/CA is extremely dense and concentrated (as Figure 2 shows) confirms the tendency toward convergence in the recent years.

This being said, the network graphs hide the fact that the authors and systems analysed here are spread over a period of more than 30 years, and that most of their work only lasted for a couple of years. Many efforts on developing dialogue-based CALL systems have not been continued after a first attempt and a few publications: teams seem to have moved on, probably because of the difficulty of developing these systems, but maybe also discouraged by the absence of a community of researchers to interact with. This phenomenon is particularly perceptible after the mid 1990s and the waning interest in tutorial CALL (Hubbard & Bradin Siskin, 2004) and, sadly, recent evolutions do not

seem to indicate a trend toward a more structured field. The lack of a research community in the field is thus strongly associated with a lack of sustainability.

Another consequence is that, to date, of all the systems we mentioned, practically none have made it to a general audience[4]. They remain at a level of proof-of-concept, or as an internal prototype. The commercial language learning programs, on the other hand, still lack of dialogue-based CALL abilities.

4 TOWARDS A TYPOLOGY OF DIALOGUE-BASED CALL SYSTEMS

Our collected set of studies refers to 96 different systems or applications[5], among which 83 were specifically designed for CALL (the other 13 have been studied by CALL researchers for their potential for language learning, but where initially created for other purposes). These systems exhibit a considerable variation of both instructional and technological designs.

Existing theoretical categorisations of dialogue systems in NLP (see Jokinen & McTear, 2010; Klüwer, 2011; Lison, 2014) distinguish systems exclusively on their internal functioning logic, focusing particularly on dialogue management, and do not account for many features that have interactional and pedagogical implications. Besides, some of the systems that fit our description of dialogue-based CALL would not spontaneously be considered as dialogue systems, because of their predetermined nature (e.g. Cucchiarini et al., 2014; Stewart & File, 2007). The typology of dialogue games developed by Piwek (2017) comes closer to our needs, but still remains directed by technological paradigms and real-world applications of dialogue systems that only partially fit dialogue-based CALL.

In consequence, we attempted to develop a typology specific to dialogue-based CALL systems. This typology was built as an iterative and bottom-up process to determine classes based on explicit criteria. First, all systems' interactional (e.g. spoken or written modality, open-ended or goal-oriented interaction, constraints on the interaction), instructional (e.g. corrective feedback) and technological features (e.g. dialogue management approach) were coded by the first author, following a coding scheme that was itself constructed iteratively throughout the analysis[6]. Second, systems were clustered around their common characteristics in various attempts to identify the most discriminant variables and, following Occam's razor, the simplest typology. While no variables allow for perfectly hermetic groups, the level of constraints imposed on the learner production seems to allow for the most consistent categorisation, as it entails directly or indirectly many instructional, interactional and technological characteristics of each type of system.

We first describe here our refined categorisation of constraints on form and meaning, and the resulting typology of dialogues and systems. We sub-

[4] Except for the general-purpose chatbots. Alelo's TLCTS may also be an exception, but it is unclear to us if their system has already been made available to language learners beyond their initial military target audience (see Johnson & Valente, 2009). It also seems that GenieTutor (Kwon et al., 2015) was available for end-users as a pilot (http://genietutor.etri.re.kr), but has since been suspended.

[5] Certain systems are analysed in various publications, such as CSIEC, which is the object of 10 papers from Jia and colleagues. At the same time, other papers analyse the potential for language learning of various (similar) systems, designed independently from the researcher (Coniam, 2008, 2014).

[6] The coding scheme for systems is appended in Annex V.

sequently identify the instructional implications, interactional characteristics and technological features for each type of system.

4.1 Constraints on learner production at the foundation of dialogue-based CALL systems

At the backbone of instructional design lies the balance between predetermined choices, or constraints, set by the system, and the degrees of freedom for the learner (Colpaert, 2004). In dialogue-based CALL, the constraints — or absence thereof — applied to the learner's production constitute indeed the primary design option and the founding criterion of our typology. These constraints are of utmost importance, as they will have direct consequences on the conversational interaction, on the potential learning outcomes, and on the required technological processing.

The learner's interactional turn can be constrained on *meaning* — the degree of negotiability of the content of each message —, or on *form* — limiting the range of linguistic items that they can use. Form and meaning are not subjected to constraints in a dichotomous way: they rather follow a continuum going from totally constrained (pre-set form/meaning) to totally unconstrained production (free form/meaning) (see also Bailey & Meurers, 2008; Wilske, 2015). To operationalise this continuum, we divided it into four levels of constraints: pre-set form/meaning, explicit constraints, implicit constraints, and no constraints[7]. Table 3 lists for every level of constraints all the examples of constraints on form and on meaning found in the reviewed dialogue-based CALL systems.

The difference between *explicit* and *implicit* constraints is important. Explicit constraints are imposed on the learner by the learning environment, externally from the communication situation. They can for example take the form of a list of words to use (explicit constraint on form), or an instruction about what message to express (explicit constraint on meaning). They are artificial, and make obvious the educational purpose of the dialogue, as they would not occur in a real-life conversation. Implicit constraints, on the other hand, are integrated, 'ecological': they originate in the communication situation, implied by the context, the previous turns (including questions asked by the system), or a certain task that should be achieved. Similar constraints occur naturally in every real-life dialogue, without anyone opposing it. The influence of implicit constraints is predominantly perceived at the meaning level. Their effect on form (e.g. formality and registers) is less marked and significant than for explicit constraints.

4.2 A typology of dialogues and systems

The combination of levels of constraints on meaning and on form constitutes the basis of our typology of dialogue-based CALL systems, along the

[7] Some would argue that no production is ever completely free of constraints. It is right, in the sense that some context always impose itself on the intention behind the expressed meaning, and that physical limitations restrain the phonemes and graphemes one can use in the form of their messages. However, for simplicity's sake, we will present as 'unconstrained' dialogues where (1) no context is explicitly set before the start of the dialogue (except for the communication window of the interface), (2) the user initiates the interaction in the way he wants to, and (3) any sequence of characters or phonemes can be uttered. This differentiates systems that offer an 'empty' communication context, abstracted from a specific reality (no constraints on meaning), from systems that place the user in a defined situation or assign him a clear task (implicit constraints on meaning).

	PRE-SET	EXPLICIT CONSTRAINTS	IMPLICIT CONSTRAINTS	NO CONSTRAINTS
CONSTRAINTS ON MEANING	No possibility to change meaning. • Text to be translated • List of utterances with unique meaning, verbatim prompt, gap-filling, rearranging words... (see EXPLICIT CONSTRAINTS ON FORM).	• Next turn of the dialogue is already set • Questions that elicit a predetermined answer • Prompts instructing what meaning to express • List of utterances (different meanings) to be chosen from.	• Set domain or context for conversation • Visual context • Task to be completed • System-initiated conversation • Questions asked by the system.	At the beginning of the dialogue, no context is set and no instructions are given to the user regarding what to say
CONSTRAINTS ON FORM	No possibility to change form. • List of utterances with different meanings (see EXPLICIT CONSTRAINTS ON MEANING)	• Limited syntax to follow (e.g., directive verb + object) • List of words/ structures to use • Words or blocks to be rearranged • Part of the utterance is already given (gap-filling) • List of utterances (same meaning) to be chosen from • Utterance to be read out loud (verbatim prompt).	[No observed case]	Free input: the user can enter/utter any sequence of characters or phonemes in the dedicated field or time frame.

Table 3 – Levels of constraints on meaning and form of the learner production, with examples for each level.

constrained-unconstrained continuum. Through the possible combinations of constraints, we identified seven instructional types of dialogue, detailed in Table 4:

1. BRANCHING DIALOGUE (2 systems): the learner has to choose ('point and click') among a list of utterances expressing different messages, all appropriate and grammatically correct. Form can thus not be modified (pre-set) and meaning is limited to the few options presented (explicit constraint).

2. FORM-FOCUSED EXERCISES (IN A DIALOGUE) (8 systems): meaning cannot be altered by the learner and form is limited to selecting, arranging, completing or translating sequences or words, or reading a given utterance out loud (for pronunciation training). The activity is performed inside a meaningful dialogue, but the focus is only on linguistic forms.

3. ELICITED DIALOGUE (12 systems): meaning and form are constrained explicitly, through a list of utterances (same meaning) selected and read out loud (e.g. Harless, Zier & Duncan, 1999), or through very precise instructions asking to express a defined meaning with a target structure (Eskenazi & Hansma, 1998).

4. MEANING-CONSTRAINED DIALOGUE (14 systems): meaning is constrained explicitly, e.g. with the system asking questions whose answers are predetermined (Kwon et al., 2015), instructing exactly what to say. The system thus defines the internal logic of the interaction.

5. FORM-CONSTRAINED DIALOGUE (5 systems): meaning is constrained implicitly, typically because the user has to interact within a very specific

context (*microworld*), such as a room, often presented visually, and at the same time, form has to follow a defined syntax (e.g. verb + object) or to use target structures.

6. CONTEXTUALISED DIALOGUE (33 systems): here, no constraint lies on form, while meaning is implicitly constraint, by a defined context (e.g. a restaurant) or task (e.g. booking a hotel room), or because the system initiates and guides the conversation, often by asking questions (whose answers are not predetermined). This is the most represented category, as implicit constraints offer a flexible framework and many creative possibilities.

7. FREE DIALOGUE (22 systems): no constraint, nor context is imposed upon the interaction, and the user is free to guide to conversation anywhere. It is usually the approach of chatbots.

	NARRATIVE S.	FORM-FOCUSED SYSTEMS			GOAL-ORIENTED SYSTEMS		REACTIVE SYS.
	BRANCHING DIALOGUE	FORM-FOCUSED EXERCISES	ELICITED DIALOGUE	MEANING-CONSTRAINED DIALOGUE	FORM-CONSTRAINED DIALOGUE	CONTEXTUALIZED DIALOGUE	FREE DIALOGUE
CONSTRAINTS ON MEANING	Explicit	Pre-set	Explicit	Explicit	Implicit	Implicit	None
CONSTRAINTS ON FORM	Pre-set	Explicit None translat.	Explicit	None	Explicit	None	None
TYPICAL CONSTRAINTS	List of utterances (distinct meanings)	Gap-filling, list of words, translation...	Branching dialogue to be uttered	Questions with predetermined answers	Target words or structure to use within a task	Task and context	No context, user guides the dialogue
# SYSTEMS	2	8	12	14	5	33	22
PUBLICATIONS	2007–2011	1986–2014	1997–2012	1992–2016	1986–2009	1982–2017	1987–2016
EXAMPLES OF SYSTEMS	LET'S CHAT (Stewart & File, 2007), MENTIRA (Holden & Sykes, 2011)	LET'S TRANSLATE (C. Wang & Seneff, 2007), de Wet et al., (2009), VIRTUAL LANGUAGE PATIENT (Walker et al., 2011), Su et al. (2013), ARTUR (Engwall et al., 2004), GREET (Cucchiarini et al., 2014)	MILT-DSR (Kaplan et al., 1998), FLUENCY (Eskenazi, 1999), CONVERSIM (Harless et al., 1999), VILTS (Rypa & Price, 1999), CANDLETALK (Chiu et al., 2007), SAYBOT (Chevalier & Cao, 2008)	MILT-TXT (Holland et al., 1999), PILÉFACE (Lelouche, 2008), CALL-SLT (Rayner et al., 2012), DIALOGUE DUNGEON (Cornillie et al., 2013), GENIETUTOR (Kwon et al., 2015)	FLAP (Culley et al., 1986), SPION (Sanders & Sanders, 1995), RECALL (Krüger & Hamilton, 1997), DREISTADT (Lech & De Smedt, 2006), TE KAITITO (Vlugter et al., 2009)	SUBARASHII (Bernstein et al., 1999), SCILL (Seneff et al., 2007), DEAL (Hjalmarsson et al., 2007), TLCTS (Johnson & Valente, 2009), SPELL (Morton et al., 2012), POMY (Lee et al., 2014)	Various chatbots in Coniam (2008, 2014), AGHATE (Williams & van Compernolle, 2009), CSIEC (Jia, 2009), VERBOT (Sha, 2009)

Table 4 – Typology of dialogue-based CALL systems, defined by the level of constraints on user production.

For simplicity purposes, these seven combinations of constraints and types of instructional dialogue can be grouped into four types of systems. In particular, certain types of dialogues are under-represented recently (e.g. only 5 form-constrained dialogues, none after 2009), while they share major characteristics with others corresponding to the same type of constraints on meaning. We thus formulate a simplified typology of systems, depending on the main objective of each:

1. NARRATIVE SYSTEMS: in a *branching dialogue*, the main objective of the interaction is to build up an emergent narrative, where the user is involved by his choices.

2. FORM-FOCUSED SYSTEMS: *form-focused exercises*, *elicited dialogues* and *meaning-constrained dialogues* share an emphasis on form, as the dialogue — being explicitly constrained on meaning — is here mostly an excuse to practice certain target structures. Within those systems, form-focused exercises can be considered as closed activities, while the other two offer a little bit more space, as half-closed activities (Desmet, 2007).

3. GOAL-ORIENTED SYSTEMS: on the contrary, *form-constrained* and *contextualised dialogues* propose to the learner to use the dialogic interaction to attain a goal (accomplishing a task or answering correctly to a set of questions), which serves as an implicit semantic constraint.

4. REACTIVE SYSTEMS: *free dialogues*, as they are not constrained by any pre-established element, leave entirely to the user the managing of the interaction, which is considered to be open-ended. The system only tries to respond in an appropriate way to every user message. The dialogue only ends if the user ceases to send messages.

Nevertheless, the value of any typology does not lie in itself, but in how it allows to understand better its object. In the following sections, we will highlight the instructional, interactional and technological implications of each kind of dialogue and system.

4.3 Instructional characteristics

The position on the continuum of constraints has direct implications, both pedagogical and interactional. For example, the input modality varies according to the type of dialogue: user production is mainly spoken in *elicited dialogues*, and written in *branching*, *form-constrained* and *free dialogues*. Considering the general trend toward multimodal interfaces, as many systems implement both spoken and written interaction (including written transcription of spoken utterances), as well as visual information (e.g. embodied agents, with facial expressions and gestures), we distinguished here the systems on their primary input modality (how the user is expressing himself). It is interesting to note that multimodal interfaces are often found within *contextualised dialogues*, as the visual information is very efficient to convey a communicational context for the interaction. Some of these systems even integrate non-verbal input opportunities, such as gestural or haptic interfaces, for the learner (e.g. Hassani et al., 2016).

The constrained-unconstrained continuum has a parallel in the continuum between controlled practice and free practice observed in language learning activities (R. Ellis, 1988). Consequently, it also has a direct implication regarding the language instruction paradigm adopted by the system. Systems can be distinguished depending on whether they bring the attention to forms in isolation (*focus on formS* (*FonFS*)), to forms as they appear incidentally in a primarily meaning-focused activity (*focus on form* (*FonF*)), or to meaning only, excepting acquisition of structures to be completely incidental (*focus on meaning* (*FonM*)) (Long & Robinson, 1998; Norris & Ortega, 2000).

Focus on form is often realized through corrective feedback provision inside a meaningful interaction (Loewen, 2011). Corrective feedback provision,

or the absence thereof, is thus also an important characteristic for dialogue-based CALL systems, as it defines whether or not the learner's attention will be brought toward form. *Form-focused systems* systematically implement explicit corrective feedback, while *narrative* and *reactive systems* never. Within *goal-oriented systems*, there is a certain variation space, depending on design choices whether or not to implement corrective feedback, and between explicit and implicit forms of feedback. Table 5 identifies the respective tendencies for each system type.

	NARRATIVE S.	FORM–FOCUSED SYSTEMS			GOAL–ORIENTED SYSTEMS		REACTIVE SYS.
	BRANCHING DIALOGUE	FORM–FOCUSED EXERCISES	ELICITED DIALOGUE	MEANING–CONSTRAINED DIALOGUE	FORM–CONSTRAINED DIALOGUE	CONTEX–TUALIZED DIALOGUE	FREE DIALOGUE
LEARNER OUTPUT MODALITY	Written	Spoken / Written	Spoken (mainly)	Written / Spoken	Written	Spoken / Written / Multimodal	Written (mainly)
FOCUS OF INSTRUCTION	FonM	FonFS	FonF	FonF	FonF / FonM	FonF / FonM	FonM
CORRECTIVE FEEDBACK	None	Explicit CF	Explicit CF	Explicit CF (mainly)	Explicit CF (mainly)	None / Implicit CF (mainly)	None
INITIATIVE	N/A	N/A	System	System	User / System	User / System	User
INTERACTIVITY	Interactive	None	None	Success / Repeat	Interactive	Interactive	Reactive
GOAL–ORIENTED INTERACTION	System-defined[a]	System-defined[a]	System-defined[a]	System-defined[a]	Goal-oriented	Goal-oriented	Open-ended

[a] As the meaning is constrained by the system initiative or instructions, the purpose of the interaction relies entirely on the system and is not always apparent.

Table 5 – Instructional and interactional features of each type of system.

Beyond the questions of focus on form and corrective feedback, the level of constraints will also impact the *task complexity*. As constraints on form and meaning decrease, the learner's attention is directed towards more and more aspects of production, augmenting the referential knowledge needed and the complexity of resource-directing variables (see Robinson, 2011). Concurrently, less constrained and more complex tasks 'promote the use of self-chosen language and thematic knowledge' (Quixal & Meurers, 2016, p. 43).

4.4 Interactional implications

4.4.1 Initiative management

The first interactional implication concerns the handling of *initiative* in the dialogue, i.e. the question of who leads the conversation, similar to the notion of (holding the) *floor* in conversation analysis (Edelsky, 1981). In a typical human conversation, initiative switches back and forth between participants, but in an artificial — and, moreover, pedagogical — interaction, initiative usually follows a default behaviour: either the dialogue is *system-initiated*, if the virtual agent leads the conversation by asking the user questions, or it is *user-initiated*, if it leaves to the learner the role of asking questions or giving orders. Naturally, to perform tasks that reach beyond simple question-and-answer series, both speakers may need to be able to sway the dialogue flow. More advanced systems thus allow for *mixed initiative*, allowing for example negotiation and collaborative planning, but even then, a certain dominance

of either the user or the system remains: we will, in such cases, talk about *user/system-initiated dialogues allowing for mixed initiative*.

In system-initiated dialogues, the virtual agent will typically use *directive prompts* ('Which station [from a certain city] would you like to depart from?', 'At what time?'), limiting the possible response to a limited set. The incorporation of *open prompts* ('How can I help you?') in such systems then asks for mixed-initiative management (see Singh, Litman, Kearns & Walker, 2002). By contrast, in user-initiated dialogues, the user is often free to guide the conversation where he pleases, and the challenge for the system consists in providing appropriate responses.

4.4.2 Goal-orientation

The second interactional implication concerns whether the dialogue is goal-oriented or not. Free dialogues, as in chatbots, are non-goal-oriented, known as *open-ended* interaction: the conversation is an end in itself, as in small talk. Being extremely difficult to foresee globally, open-ended dialogues tend to be strictly user-initiated and reactive, only managed at the turn-level, as adjacency pairs (e.g. salutation–salutation, compliment–thanks, question–answer). As there is no preset objective, the conversation tends to continue as long as the user keeps taking their turn.

By contrast, *goal-oriented* interactions seek the accomplishment of a task (in the dialogue systems literature, they are often referred to as *task-oriented*). The tasks to perform vary widely, from everyday transactions (e.g. buying something, asking for directions, booking a train ticket, consulting a doctor) to professional transactions (e.g. job interview, group meeting, deal negotiation). This approach of dialogue is consistent with the idea of conversation as a 'joint activity' where people try to collaboratively attain common goals (Clark, 1996). Even though they require more complex dialogue models, goal-oriented interactions tend to be more predictable, as they follow certain patterns depending on the task. Besides, from a pedagogical point of view, they correspond to a task-based approach to language learning and teaching (Long, 2015).

4.5 Technological implications

4.5.1 Variation, predictability and processing

Our typology has also technological implications. As the constraints decrease, the potential variation in learner production augments, and its predictability plunges (Desmet, 2006). As a result, the complexity of its automated analysis increases exponentially, the number of possible combinations becoming quickly impossible to foresee (Bailey & Meurers, 2008). Such exponential variation not only makes it difficult to provide corrective feedback, but especially complicates the dialogue management and response generation. For this reason, many dialogue-based CALL systems have avoided complex learner language processing with constraints on meaning and on form: multiple choice of words or utterances, gap-filling, or blocks rearranging exercises do not require natural language understanding techniques. The only processing needed is for error diagnosis and feedback, which for very short free input strings, can be rather simple (Wilske, 2015). These implications for each type of dialogue are described in Table 6.

The ability to *process meaning* (natural language understanding) is thus a key differentiator between explicitly and implicitly constrained dialogue-

	NARRATIVE S.	FORM-FOCUSED SYSTEMS			GOAL-ORIENTED SYSTEMS		REACTIVE SYS.
	BRANCHING DIALOGUE	FORM-FOCUSED EXERCISES	ELICITED DIALOGUE	MEANING-CONSTRAINED DIALOGUE	FORM-CONSTRAINED DIALOGUE	CONTEXTUALIZED DIALOGUE	FREE DIALOGUE
FORESEEN VARIATION	None	−−	+	+	++	++	+ + +
PREDICTABILITY	[+ + +] (fully predictable)	+ + +	++	++	+	+	−
MEANING PROCESSING	As dialogue branch (fully predictable)	No (fixed)	No (assumes expected meaning)	Yes (validation of expected meaning)	Yes (limited scope)	Yes	Practically impossible → Avoidance strategies
DIALOGUE CONTROL	Graph	Fixed path	Fixed path	Fixed path	Frame / Graph	Frame / Probabilistic	Rules-based matching
LEVEL OF PROCESSING FOR DIALOGUE CONTROL	No processing (state only)	No dialogue control (fixed)	No dialogue control (fixed)	No dialogue control (fixed)	Basic semantic representation (e.g. ⟨`action + object`⟩)	Semantic representation (e.g. dialogue act)	Keywords / patterns in normalized utterance

Table 6 – Technological characteristics and tendencies of each type of dialogue-based CALL system.

based CALL systems. Such ability is often considered as a requirement for a communicative approach to language learning (Amaral & Meurers, 2011). Yet, due to the complexity of taking into account and acting upon an indefinite number of possible meanings from user's utterances, many systems across the years have eluded the problem. Form-focused and elicited dialogues avoid meaning processing, assuming the learner produced the elicited meaning. Meaning-constrained dialogues tend to only validate if the expected meaning was produced. At the other end of the continuum, chatbots in free dialogues are supposed to react upon meaning, but because the search space is so broad, they tend to use avoidance strategies, such as setting up catch-all fallback responses, giving vague answers or returning the question. Implicitly constraining the possible content of learners' messages is thus central, and designing conversational tasks that are both meaningful and reliably analysable by NLP strategies is the main challenge of dialogue-based CALL (Wilske, 2015).

4.5.2 Dialogue management and natural language understanding

Finally, a key issue of dialogue systems is the *management* of the dialogue flow, especially the *control* of the agent response to the user. Of course, if only one conversational path is possible, either because the learner cannot affect the content of the message (form-focused), or because the system will only accept one excepted meaning (elicited or meaning-constrained), then no processing is required: the conversation follows a fixed path. No processing either is required for branching dialogues: they follow a graph defining all the possible conversational paths, and the next turn is triggered directly by the user action (utterance selection). In contrast, in the implicitly constrained and unconstrained dialogues, dialogue management is a challenge, which has been addressed in many different ways by chatbots and dialogue systems. It is beyond the scope of this chapter to enter into technical details (for a review of dialogue management in NLP, see Jokinen & McTear, 2010; Jurafsky & Martin, 2008; Lison, 2014), but general lines can be drawn.

The dialogue control decision is based on the processing of the user production, which can occur at different levels, from surface forms to deep semantic representations. Chatbots, often built with AIML (Wallace, 2003), tend to process the user production at a shallow level, by matching keywords or characters patterns (e.g. ⟨`I'd like *`⟩, where * is a wildcard allowing any continuation) in the input, after it has been normalized, in particular after "simplifying" recursively certain variations (e.g. ⟨`Can I ask you where... is please`⟩ is reduced into ⟨`Where is...`⟩). Beyond this update mechanism, the dialogue management is based on simple, deterministic matching, defined by multitudinous handcrafted rules (Klüwer, 2011).

Form-constrained and contextualized dialogue systems, on the other hand, tend to process the user production at the semantic level, even though some systems also resort to numerous matching rules. Ideally, the user utterance is normalized (e.g. errors and typos correction, capitalization and punctuation normalization...), and then analysed into a semantic representation, which can be a dialogue act representation, an intent with or without entities, or an abstract ⟨`action + object`⟩ formalism. This semantic analysis can be rules-based, usually by matching patterns into intents and entities, or can be, in more recent systems, learned from annotated data (existing corpus) in a probabilistic fashion (S. Young, 2000; S. Young, Gašić, Thomson & Williams, 2013). For dialogue management, these systems can also take into account dialogue state information, based on previous turns and inferred interactional goal, and other sources and levels of information (multimodal input, syntactic structure...).

Subsequently, depending on the representation of the dialogue state, the virtual agent response can be selected, either through a graph- or a frame-based approach, if dialogue management is rules-based and deterministic, or through statistical models such as partially observable Markov decision processes (POMDP) (e.g. S. Young et al., 2013), Bayesian networks (Lison, 2015), or recurrent neural networks (e.g. Lison & Bibauw, 2017; Vinyals & Le, 2015), when it is data-driven and probabilistic. Most dialogue-based CALL systems have actually adopted ad hoc processes, mostly handcrafted, with a reduced complexity, and there are hence ample opportunities to apply the advances made in dialogue systems to their CALL adaptation.

4.6 Summary: types and constraints as design choices

The constraints continuum, and especially between explicit and implicit constraints on meaning, might give the false impression that the ideal instructional design for dialogue-based CALL would be unconstrained. It is however far from the case.

Constraints have interactional, pedagogical and technological justifications. Artists know how constraints can promote creativity, by forcing the subject to find novel solutions or by taking them off the beaten track (Stokes, 2005). Similarly, in language learning activities, constraints can be used to counteract the creative paralysis that can occur when facing too extensive choice (Tin, 2012). They also contribute to direct the learner's attention and resources toward target language elements, and, by increasing task complexity, to promote L2 development (Robinson, 2011). From an interactional perspective, implicit constraints are naturally present in real-life dialogues: no conversation is ever initiated without a context, and many interactions have an implicit goal that structures them. Finally, from an NLP angle, considering the difficulty for ICALL systems to process learner language with a very high precision, con-

straints allow to limit and guide the potential variation that has to be processed (Amaral & Meurers, 2011).

Nevertheless, a certain level of freedom for learner production is also important. By allowing the learner to express their own personal meanings (R. Ellis, 2005), the activity is more motivating and the primary focus is brought on meaning rather than forms. The dialogue also gains in authenticity and realism. This idea was already put forward by R. Young (1988, p. 64): 'computer programs which allow outcomes to be negotiated generate the kinds of conversational discourse which are most useful for successful language learning'. This principle ensures the interactivity of the dialogue, and the feeling for the learner that their conversational actions matter. When the constraints on meaning enforce a single conversational path, negotiation of meaning and other interactional phenomena that we know are beneficial for language learning (Mackey & Goo, 2007) cannot occur.

Having said that, both constrained and unconstrained ends of the continuum are not necessarily beneficial. Form-focused exercises, on one side, because of their strong pedagogical and interactional limitations, should not be the main objective of dialogue-based CALL. On the other side, contrary to a tendency to overestimate the prevalence of small talk, open-ended out-of-context interactions are very artificial, and free conversations with reactive systems such as chatbots tend to tail off quickly, as the user has no reason to keep the conversation going. For these reasons, *free dialogue* should not be seen as the ultimate target of dialogue-based CALL either. The most interesting affordances of dialogue-based CALL seem to rely on innovative use of (mostly) implicit constraints on meaning that allow both self-expression, meaningfulness and a certain level of predictability.

5 A SYNTHESIS OF EFFECTIVENESS STUDIES' RESULTS

5.1 Types of studies

As an emerging research object, publications on dialogue-based CALL have progressively shifted from descriptions of instructional and technological design towards technical, observational and experimental evaluations of the effectiveness of systems on learning goals. We analysed and coded our corpus of publications to quantify this evolution. In order to clearly characterise the scope of each work, we distinguished six different types of studies:

(a) REVIEWS, which, remaining on a theoretical level, look into the potentialities or the reported effectiveness of systems;

(b) SYSTEM DESCRIPTIONS, which typically present an application concept or a technological component, but without any evaluation of it, nor any test by end-users;

(c) TECHNICAL EVALUATIONS, which add to the earlier a technical evaluation of the system's processing accuracy or performance (e.g. recognition rate, word error rate);

(d) OBSERVATIONAL STUDIES, which describe the use of the system by a sample of users, e.g. by analysing their linguistic productions;

(e) SURVEY STUDIES, which evaluate users' attitude or perception toward the system — typically, some quantitative post-use measures related to user satisfaction such as the Technology Acceptance Model (Davis, 1989);

(f) EFFECTIVENESS STUDIES, which measure the effects of the use of the system on some dependent variable(s), either attitudinal (e.g. motivation) or related to a language learning outcome (e.g. proficiency, accuracy, or knowledge measures), requiring at least two measure points (pre-post and/or control-experimental design).

In summary, the first two types represent theoretical research, whereas the other four involve empirical studies, with either a technical (c) or pedagogical (d, e, f) evaluation. Considering that most papers include a literature review, a system description and an empirical component, this classification should be seen as cumulative (i.e. except for reviews, all studies are likely to also include some of the previous types; e.g. a survey study might also include a brief review, a system description and some observational data).

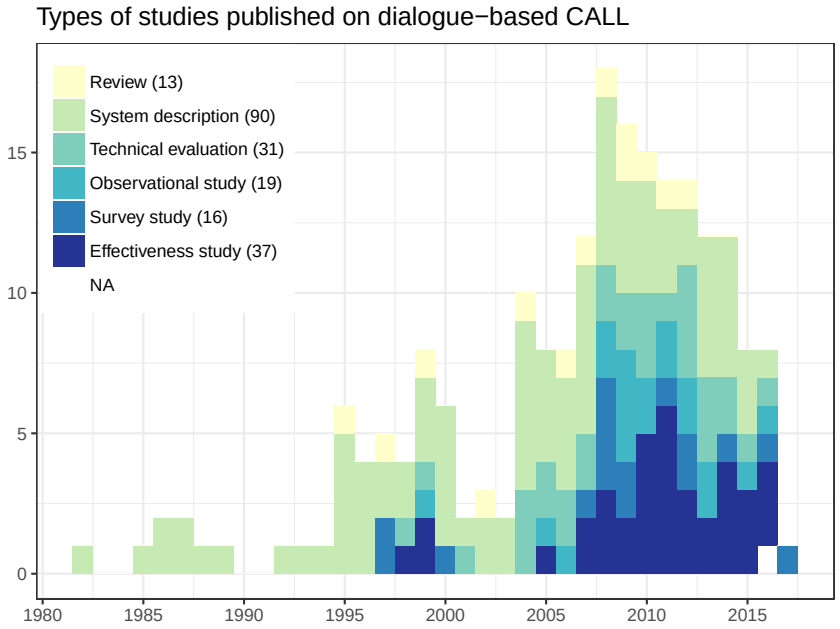

Figure 4 – Evolution towards empirical and experimental studies in dialogue-based CALL.

As Figure 4 shows, there has been a progressive shift from papers mostly describing systems (1982-2007) to technical evaluation studies (mainly in NLP) and survey studies (common in educational technology), and most recently to effectiveness studies (since 2007). This is explained both by the maturation and methodological tendencies of CALL research (Norris, Ross & Schoonen, 2015), and by the emergence of dialogue-based CALL as a research object.

To be able to establish cause-effect relations, the experimental design is the unrivalled method. However, the complexity of coordinating such experiments in learning contexts made them relatively rare in CALL research until recently (Felix, 2005). As shown in Table 7, our corpus only contains 37 papers qualifying as effectiveness studies, but, among them, research designs and methodological quality vary widely. While they all address the question of effectiveness, studies compare systems against various reference groups (e.g. face-to-face interactions, variations of the same system, or the absence of

intervention), or as a within-subjects study only, and use very different instruments to measure the outcomes, with more or less standardisation and rigour. Besides, various studies only report quantitative results partially (e.g. means without standard deviations, or graphs and p-values without reporting precise summary statistics)[8]. These facts all concur in limiting the comparability of the effects, and the generalisability of the findings.

OUTCOME MEASURE	STUDY TYPE			TOTAL
	OBS.	SURVEY	EFFECT.	
User engagement/system usage	5	5	3	13
User attitude toward system	8	16	15	39
In-system L2 performance characteristics	8	4	5	17
Effectiveness on motivation	1		8	9
Effectiveness on L2 development	3		30	33
Productive skills	3		28	31
by mode:				
Speaking	2		20	22
Writing	1		10	11
by measured construct:				
Holistic proficiency	1		10	11
Complexity (incl. vocabulary)	1		10	11
Accuracy	2		20	22
Fluency			7	7
Receptive skills			4	4
Other L2 development measures			3	3
TOTAL NUMBER OF PUBLICATIONS	19	16	37	72

A publication may report on results from various tests and measurements, thus vertical values are not mutually exclusive.
Obs. = Observational study. Effect. = Effectiveness study.

Table 7 – Number of papers by category of study and outcome measure.

We will try to summarise hereafter the conclusions from the empirical pedagogical evaluation studies[9] on the effectiveness of dialogue-based CALL.

5.2 Observational studies

The observational studies in our set of publications have, for most of them, essentially an exploratory value, of utility to identify strengths and weaknesses of certain system designs. Most of the findings of such studies focus on the elements that generated interest among learners, and on the limitations or problems that occurred. Various observational studies on chatbots showed that users appreciated the freedom and the light-hearted conversations, but were also frequently interrupted by communication breakdowns and irrelevant

[8] We tried to contact the authors of the papers where some data was missing, but for the majority of them, we could not obtain the missing pieces of information. The fact that some of those studies are already about ten years old and that, as we mentioned earlier, many teams and specialists have now moved on to other research topics, can explain in part the unavailability of these datasets today.

[9] We chose not to include in this research synthesis a summary of technical (NLP) evaluation studies on dialogue-based CALL, as these studies address very different research questions (e.g. regarding the accuracy of the agent's responses), which would have raised methodological and technical issues that are beyond the scope of this review.

responses, which affected their engagement in the dialogue (Fryer & Nakao, 2009; Williams & van Compernolle, 2009). It should be noted that these miscommunication events, while present in all systems due to the limitations of current NLP technology, seem to be much more systematic in open-ended, free dialogues than in contextualized and more constrained ones, and that recognition errors in the latter are less a hindrance for users (Holland et al., 1999; Morton et al., 2008).

On a more instructional design perspective, it appeared important to provide more scaffolding and support mechanisms to the learners, to help them in the receptive and productive phases, as well as more progressive, slow paced and accessible segments for beginners (Rypa & Price, 1999; Walker, Trofimovich, Cedergren & Gatbonton, 2011).

A recent study, comparing an oral role playing activity (between a learner and a native speaker) with an application letting the students answer orally after a video prompt, also provides an interesting insight: the dialogue-based CALL users were paying stronger attention to form, even with spoken interactions (Sydorenko, 2015). The researcher observed more signs of uptake in the computer program, as learners were incorporating sequences and strategies from the input into their speech, and accuracy was improving across iterations.

5.3 Survey studies

The survey studies all analysed users' attitudes towards the system, through post-use questionnaires, and typically investigated the system's perceived ease-of-use and perceived usefulness, as these two variables have been popularised by the Technology Acceptance Model (Davis, 1989). The main assumption is that perceived ease-of-use and perceived usefulness could be the principal predictors of behavioural intention to use and, by extension, of system usage (Venkatesh & Davis, 2000). Since many dialogue-based CALL applications were conceived for self-directed learning, their effectiveness relies on the learner's willingness to use them and their acceptance is thus a sine qua non.

User evaluation results were globally positive for all systems (e.g. Ehsani, Bernstein, Najmi & Todic, 1997; Harless et al., 1999; Kaplan, Sabol, Wisher & Seidel, 1998; Schoelles & Hamburger, 1997; Sha, 2009). However, the post-use measures of user perception are usually presented without any comparison or reference point, which hinders their significance. Moreover, most studies in our corpus did not follow a standardised methodology: each used different items and scales, with their own operationalisation. Results are therefore not comparable. In addition, many papers only report the central tendency (e.g. mean) of the results, without any information on their dispersion (e.g. standard deviation), which limits their statistical value.

More precise findings tend to indicate that dialogue-based CALL systems 'can support high levels of user acceptability and engagement' (Anderson et al., 2008, p. 613). Added to a virtual world for instance, they facilitated the engagement of learners (Y. F. Wang, Petrina & Feng, 2017). When offered, gamification elements inside the system were appreciated (Baur, Rayner & Tsourakis, 2015).

Attitudes toward the program vary between learners though. University students tend to engage more and perceive a higher utility for such systems than high school or middle school students (Ehsani, Bernstein & Najmi, 2000; Jia & Chen, 2009). Two studies also showed that intrinsically motivated learners engaged and benefited more from the software (Anderson et al., 2008; Baur et al., 2015).

It is also important to note that, when offered both form-focused structured activities and meaning-focused dialogue activities, the adult learners using TLCTS considered the more structured activities to be more useful and more engaging (Surface, Dierdorff & Watson, 2007). This finding supports a vision of dialogue-based CALL as a supplement of other learning activities, and not as a stand-alone learning environment.

5.4 Effectiveness studies

5.4.1 Effects on motivation

A few papers presented results from experimental studies on the motivational effects of dialogue-based CALL, i.e. how using the system affected the learners' motivation for language learning. A common assumption, as mentioned earlier (see 1.2), is that the ability to practice the L2 in a meaningful and realistic setting might improve the user's L2 self-confidence, and thus his willingness to communicate (MacIntyre, Dörnyei, Clément & Noels, 1998), as well as giving a more tangible image of the communicative goals guiding the language learning process.

A couple of studies demonstrated that the users' self-efficacy, i.e. their confidence in their ability to learn, speak and understand the target language significantly increased as a result of the system use (Surface et al., 2007; N. Wang & Johnson, 2008). Another team, working on the SPELL system, observed a similar motivation boost with two groups of university students, who reported being more relaxed interacting with the virtual agents than with a human (Anderson et al., 2008). Studying specifically the impact on willingness to communicate, Ayedoun et al. (2015) observed a clear rise in confidence and a slightly lower nervousness among students after using the dialogue-based CALL application.

While the previous studies focused on adult learners, S. Lee et al. (2011) evaluated the impact of introducing two conversational (physical) robots for English learning in a South Korean elementary classroom during 8 weeks. Their paper is also of particular interest because of its rigorous experimental design, its thorough validation of instruments and its detailed reporting of methodology and results. Their within-subject comparison (repeated measures design) reveals significant effects of the system on the three motivational variables they considered — interest in learning English ($d = .59$), confidence with English ($d = .66$), and motivation for learning English ($d = .98$). Compared with domain-specific reference values for effect sizes (Plonsky & Oswald, 2014), these values correspond to a small (but clearly significant) effect size on interest and confidence, and a medium effect size on motivation.

Naturally, those findings are too scattered to confirm the initial hypothesis. They should be established for several kinds of system, and with different operationalisations of L2 learning motivation. Yet, they show a consistent trend among studies, and pave the way for future research.

5.4.2 Effects on language development

The principal declared objective of every dialogue-based CALL system is to help learners develop their L2 proficiency. It is hence unsurprising for their effectiveness on this matter to be the primary target of evaluation. As shown in Table 7, 30 papers reported experimental evaluations of a certain system on different L2 learning outcomes. The vast majority focused on PRODUCTIVE SKILLS,

especially speaking proficiency, that were measured either through some of their dimensions, i.e. complexity, accuracy and fluency, or through a holistic evaluation. The choice of productive skills, especially speaking, for evaluation confirms the shared assumption among dialogue-based CALL researchers that its major impact should be on the procedural aspects of language learning.

As with the other groups, not all papers report enough data to be quantitatively comparable and construct a summative perspective of the effectiveness of such systems. Yet, all present interesting insights and encouraging results regarding effectiveness. One of the most compelling findings is reported by Vlugter et al. (2009), who compared the effectiveness of their written dialogue system (Te Kaitito) with both a control group, and a group who received similar interactions from a human tutor (as the 'gold standard'). Their evaluation on the acquisition of pronouns in Māori reveals that the virtual tutor was as effective as the human tutor, and significantly outperformed the control group (Vlugter et al., 2009). After setting his CSIEC chatbot in a middle school classroom for six months, Jia et al. (2013) observed major improvements on exam scores (+27%) and a critical shift for the treatment class, which overtook the other groups. The Tactical Iraqi version of TLCTS also increased significantly ($p < .01$) both linguistic and cultural knowledge of the trainees, with clear improvement of their oral proficiency (Surface et al., 2007). Such positive effects occur similarly with form-focused systems, such as the CAPT application for Chinese pronunciation developed by Su, Wu and Lee (2015), which fostered high improvements in individual phoneme pronunciation.

Regarding the CONTEXT OF USE, most studies were organized in blended environments, where dialogue-based CALL was implemented as a supplement to classroom instruction. This blended design is supported by research findings, among others on TLCTS, which was one of the few systems that have been tested as the sole mode of instruction. It shows that, in most cases, participants preferred the dialogue-based application in addition to normal training courses (Surface et al., 2007). The self-directed learning with TLCTS alone was only recommended for highly responsible learners, within a structured environment that would provide high levels of guidance and feedback. Beyond the question of blended vs. independent use, general findings do not seem to identify significant differences between learning context, or between different age groups. For example, Jia (2009) saw no perceptible difference between middle school and graduate students in the effectiveness of CSIEC. Results in elementary schools are consistent with the general effectiveness trend (S. Lee, Noh, Lee, Lee & Lee, 2010).

On the other hand, the effects of the intervention seem to be greater for learners with a low to moderate proficiency beforehand (Kaplan et al., 1998). Probably because there was more room for improvement, and because of a certain ceiling effect for advanced learners, Chiu, Liou and Yeh (2007) observed a higher increase in proficiency among intermediate learners than among English-major students. N.-Y. Kim (2016) also observes a significantly higher effect on oral proficiency with beginners than with intermediate and advanced learners.

Until now, we have analysed the different studies as if the systems were identical, but we have seen that they vary widely on many levels (see our typology in 4.2). Because of the absence of a global point of view of dialogue-based CALL, few studies have compared its variations. Yet a few INSTRUCTIONAL FEATURES have been the subject of effectiveness studies. For instance, Suzuki, Nose, Hiroi and Ito (2014) observed that adding an expression of time pressure,

through the embodied agent, reduced significantly the users' switching pause duration, in particular limiting its dispersion.

Wilske and Wolska (2011; Wilske, 2015) dedicated various studies at studying the impact of CORRECTIVE FEEDBACK and form-focused instruction in dialogue activities. Their findings on dialogue-based CALL are coherent with the literature on feedback and focus-on-form in SLA: intentional and explicit feedback in form-focused practice, as well as incidental, implicit focus-on-form in a primarily meaning-based task, both contributed to the acquisition of a greater accuracy in the target structures, but explicit feedback had a stronger impact than incidental feedback (Wilske & Wolska, 2011). Petersen (2010) also established that the provision of implicit corrective feedback (recasts) had equally positive effects on language development in a written computer-guided interaction as in a face-to-face oral conversation with a native speaker.

The general impact of dialogue-based CALL systems on language learning in all these effectiveness studies is almost always positive, independently from the outcome measures considered. All systems seem to boost both speaking and writing skills, measured holistically or through specific components (complexity, accuracy, fluency, vocabulary). However, in a majority of studies, the measured effects fail to achieve significance, most often because of insufficient sample sizes. Besides, a serious limitation also arises from the fact that, in most cases, researchers evaluate the effectiveness of a system they contributed to create, rising the problem of a potential confirmation bias in their studies. For all these reasons, a definitive verdict on the effectiveness of dialogue-based CALL on L2 development is still difficult to establish, and more experimental research, with rigorous designs and evaluation instruments, is seriously needed.

6 CONCLUSIONS AND AVENUES FOR RESEARCH

Arising as different initiatives in the ICALL, computer-assisted pronunciation training, spoken dialogue systems/conversational agents, and chatbots traditions, dialogue-based CALL has progressively constituted itself as an emerging research domain. Although it still lacks a certain self-awareness, sustainability and publicity in its undertakings, major progresses have been achieved in the last 10 years from bringing together techniques from NLP and spoken dialogue systems, instructional design expertise from ICALL and evaluation methods from SLA.

We proposed an inclusive definition of dialogue-based CALL as any system allowing a user to have a dialogic interaction with an automated agent as a language learning task. The field has much to gain from combining efforts currently led under many denominations, from personal assistants and bots to automated conversations in virtual worlds and games. Particularly, the current public enthusiasm for artificial intelligence and systems such as SIRI and ALEXA will certainly give rise to research interests from new researchers: it is crucial that their work articulates with the existing body of literature on dialogue-based CALL, rather than appearing as disconnected research items on a seemingly 'new technology'.

This larger and more coherent perception of the field is also essential to strengthen the research community around it. Being part of a more global effort might help research teams sustain their focus on conversational applications, and bring their applications to a public release state, rather than as internal prototypes. This is crucial to allow for external evaluations and

comparisons of systems. Very recent joint research efforts from academia and industry are showing, on this aspect, promising results (Sydorenko, Smits, Evanini & Ramanarayanan, 2018).

We put forward a systematic typology of dialogue-based CALL systems and instructional dialogues based on explicit categorization criteria, which materializes into seven types of dialogue and four major types of systems: *narrative*, *form-focused*, *goal-oriented* and *reactive systems*. We described how this typology had instructional, interactional and technological implications on both the design and educational potential of each system, the constrained-unconstrained continuum having an echo in the shift between focus on forms, focus on form and focus on meaning. To design a dialogue-based CALL application is to find an adequate balance between constraints, which guide and focus the user production, to reduce its unpredictability and allow its automated processing, and freedom left to the learner to express their own meanings interactively.

While there have been syntheses and typologies of ICALL (e.g. Gamper & Knapp, 2002) or general-purpose dialogue systems (in NLP), this is the first attempt at categorising conversational systems for language learning. This typology was built in a systematic data-driven approach, identifying classes based on explicit criteria. Its value also resides in the effort of linking together instructional decisions, interactional behaviours and technological options.

The focus on the combination of form and meaning constraints is essential in analysing previous systems, but also to envision future ones. Our synthesis shows that the most promising approach for future systems development involves disruptive forms of implicit constraints on meaning, in order to afford the self-expression of intents by the learner while still being partially anticipated by the system, through task-based interactions with some level of mixed initiative. In other words, *goal-oriented systems* with *contextualised dialogues*. To ensure higher effect on learning outcomes, systems should also implement some form of corrective feedback and provide scaffolding to support the learners in their production.

Nevertheless, many questions of instructional and technological design remain open. What are the most effective approaches to dialogue management and natural language understanding? How to efficiently build up dialogic content for these systems, to allow them to hold longer and more diverse conversations? How can systems nurture the motivation of users and make sure they remain engaged on the long-term? A dialogue-based CALL — as any other CALL — system's effectiveness lies entirely in how and how much it is used by the learners (Surface et al., 2007). As studies have shown that it provides the best results in a blended learning setting, questions remain of how to use it in complement with a regular language instruction, and teachers still face challenges when trying to integrate dialogue-based CALL into their classrooms or online courses.

The last section summarised results of empirical studies on the effectiveness of dialogue-based CALL on attitudinal and cognitive variables. Regarding context and population, such systems have been used with more engagement by university students (more than to younger learners) and intrinsically motivated learners, and have had stronger impact on beginners and lower intermediate learners. Regarding motivation, various independent studies demonstrated that dialogue-based CALL has a significant effect on various aspects of learner's L2 motivation. In particular, it is effective in raising their self-confidence and lowering their anxiety, thus positively affecting their willingness to communicate.

The effects on language learning outcomes are generally positive, and a study even demonstrated an effectiveness similar to a human tutor. Research has shown that learner's attention to form is raised, thus possibly improving accuracy. Generally, the strongest effects have been observed on oral proficiency.

However, while positive effects have been presented in various studies, they do not always achieve significance, nor do they always provide comparable measurements. This is often due to small samples ($n \leq 20$) and short interventions (1-3 sessions, with a total time on task rarely exceeding 2 hours). Robust experimental studies on dialogue-based CALL are still scarce, and we need more systematic results to be able to draw firm conclusions on the question of its effectiveness. Future research wanting to establish quantifiable evidence on dialogue-based CALL effectiveness should use standardize instruments measuring specific dimensions of L2 proficiency, or the acquisition of specific target structures, and ensure the sample size and duration of their intervention provide enough power to test their hypotheses.

Future research on dialogue-based CALL can take two complementary roads. The first one is to explore further the relative effectiveness of certain design features and components of the systems themselves. Such efforts should be in line with a coherent perspective of dialogue-based CALL research, taking into account what was already established. Observational research can better document the actual use of general purpose conversational systems by autonomous learners, and study potential instructional implementations of these tools into an actual course. On the general efficacy of dialogue-based CALL, we still need confirmatory evidence and more precise quantitative measures of its impact (e.g. relative effects depending on time on task). Research can also address many comparison questions on the relative effectiveness of certain design features (e.g. gamification, embodiment of agent), on different populations (e.g. age, context) and on distinct learning outcomes, such as sub-dimensions of proficiency (complexity, accuracy, fluency).

The second avenue for research is to consider dialogue-based CALL as tool to investigate language acquisition in general, and cognitive-interactionist theories of SLA in particular. As me mentioned previously (see 1.2), dialogue-based applications offer controllable forms of conversational interaction. They provide an ecologically valid setting to test hypotheses on interaction by allowing a fine-grained control over many parameters of the interaction (Cornillie, Van den Noortgate, Van den Branden & Desmet, 2017). They even provide precisely reproducible interaction, both within a single study as for replication purposes.

ACKNOWLEDGEMENTS

The authors would like to thank Detmar Meurers and Frederik Cornillie for their insightful comments on earlier versions of the manuscript, as well as the editor and the three anonymous reviewers for their valuable comments and suggestions that greatly contributed to improve this article. Thanks also to all the authors who sent us their publications when they were not available online. This research was made possible by a SENESCYT (Ecuador) doctoral scholarship to the first author.

REFERENCES

Amaral, L. & Meurers, D. (2011). On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, *23*(1), 4–24. doi:10.1017/S0958344010000261

Anderson, J. N., Davidson, N., Morton, H. & Jack, M. A. (2008). Language learning with interactive virtual agent scenarios and speech recognition: Lessons learned. *Computer Animation and Virtual Worlds*, *19*(5), 605–619. doi:10.1002/cav.265

Ayedoun, E., Hayashi, Y. & Seta, K. (2015). A conversational agent to encourage willingness to communicate in the context of English as a foreign language. *Procedia Computer Science*, *60*, 1433–1442. doi:10.1016/j.procs.2015.08.219

Bailey, S. & Meurers, D. (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 107–115). Columbus, Ohio: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology-new/W08-0913

Baur, C., Rayner, E. & Tsourakis, N. (2015). What motivates students to use online CALL systems? A case study. In *Proceedings of the 9th International Technology, Education and Development Conference* (*INTED 2015*) (pp. 2331–2338). Madrid: IATED. Retrieved from http://archive-ouverte.unige.ch/unige:46561

Bergmann, K. & Macedonia, M. (2013). A virtual agent as vocabulary trainer: Iconic gestures help to improve learners' memory performance. In R. Aylett, B. Krenn, C. Pelachaud & H. Shimodaira (Eds.), *Intelligent Virtual Agents* (pp. 139–148). Lecture Notes in Computer Science. doi:10.1007/978-3-642-40415-3_12

Bernstein, J., Najmi, A. & Ehsani, F. (1999). Subarashii: Encounters in Japanese spoken language education. *CALICO journal*, *16*(3), 361–384. doi:10.1558/cj.v16i3.361-384

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). doi:10.14705/rpnet.2015.000310

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H. & Yan, H. (2000). Conversation as a system framework: Designing embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost & E. F. Churchill (Eds.), *Embodied conversational agents* (pp. 29–63). Cambridge, MA: MIT Press.

Chen, L. & Tokuda, N. (2003). A new template-template-enhanced ICALL system for a second language composition course. *CALICO Journal*, *20*(3), 561–578. doi:10.1558/cj.v20i3.561-578

Chevalier, S. & Cao, Z. (2008). Application and evaluation of speech technologies in language learning: Experiments with the Saybot player. In *Proceedings of the Annual Conference of the International Speech Communication Association* (*INTERSPEECH 2008*) (pp. 2811–2814). Brisbane, Australia: IEEE. Retrieved from http://sylvainchevalier.free.fr/Publications/Chevalier-Interspeech2008.pdf

Chiu, T.-L., Liou, H.-C. & Yeh, Y. (2007). A study of web-based oral activities enhanced by automatic speech recognition for EFL college learning. *Computer Assisted Language Learning*, *20*(3), 209–233. doi:10.1080/09588220701489374

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Colpaert, J. (2004). *Design of online interactive language courseware: Conceptualization, specification and prototyping: Research into the impact of linguistic-didactic functionality on software architecture* (Doctoral dissertation, Universiteit Antwerpen, Antwerpen). Retrieved from http://anet.be/record/opacirua/c:irua:48805

Coniam, D. (2008). Evaluating the language resources of chatbots for their potential in English as a second language. *ReCALL*, *20*(1), 98–116. doi:10.1017/s0958344008000815

Coniam, D. (2014). The linguistic accuracy of chatbots: Usability from an ESL perspective. *Text & Talk*, *34*(5), 545–567. doi:10.1515/text-2014-0018

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). San Marcos, TX: CALICO.

Cornillie, F., Van den Noortgate, W., Van den Branden, K. & Desmet, P. (2017). Examining focused L2 practice: From in vitro to in vivo. *Language Learning & Technology*, *21*(1), 121–145. doi:10125/44598

Cucchiarini, C., Bodnar, S., Penning de Vries, B., van Hout, R. & Strik, H. (2014). ASR-based CALL systems and learner speech data: New resources and opportunities for research and development in second language learning. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 2708–2714). Reykjavik, Iceland: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/index.html

Culley, G., Mulford, G. & Milbury-Steen, J. (1986). A foreign-language adventure game: Progress report on an application of AI to language instruction. *CALICO Journal*, *4*(2), 69–87. doi:10.1558/cj.v4i2.69-87

da Costa Pinho, I., Epstein, D., Reategui, E. B., Corrêa, Y. & Polonia, E. (2013). The use of text mining to build a pedagogical agent capable of mediating synchronous online discussions in the context of foreign language learning. In *2013 IEEE Frontiers in Education Conference (FIE)* (pp. 393–399). 2013 IEEE Frontiers in Education Conference. doi:10.1109/FIE.2013.6684853

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. doi:10.2307/249008. JSTOR: 249008

de Wet, F., Van der Walt, C. & Niesler, T. R. (2009). Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication*, *51*(10), 864–874. doi:10.1016/j.specom.2009.03.002

DeKeyser, R. M. (Ed.). (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge; New York: Cambridge University Press.

DeSmedt, W. H. (1995). Herr Kommissar: An ICALL conversation simulator for intermediate German. In V. M. Holland, J. D. Kaplan & M. R. Sams (Eds.), *Intelligent Language Tutors: Theory shaping technology* (pp. 153–173). Mahwah, NJ: Lawrence Erlbaum.

Desmet, P. (2006). L'enseignement/apprentissage des langues à l'ère du numérique: Tendances récentes et défis. *Revue française de linguistique appliquée*, *11*(1), 119–138. Retrieved from http://www.cairn.info/revue-francaise-de-linguistique-appliquee-2006-1-page-119.htm

Desmet, P. (2007). L'apport des TIC à la mise en place d'un dispositif d'apprentissage des langues centré sur l'apprenant. *ITL - International Journal of Applied Linguistics*, *154*, 91–110. doi:10.2143/ITL.154.0.2023980

Edelsky, C. (1981). Who's got the floor? *Language in Society*, *10*(3), 383–421. doi:10.1017/S004740450000885X

Ehsani, F., Bernstein, J. & Najmi, A. (2000). An interactive dialog system for learning Japanese. *Speech Communication*, *30*(2-3), 167–177. doi:10.1016/S0167-6393(99)00042-4

Ehsani, F., Bernstein, J., Najmi, A. & Todic, O. (1997). Subarashii: Japanese interactive spoken language education. In G. Kokkinakis, N. Fakotakis & E. Dermatas (Eds.), *Proceedings of the Fifth European Conference on Speech Communication and Technology (EUROSPEECH 1997)*, Rhodes, Greece: ICSA. Retrieved from http://www.isca-speech.org/archive/eurospeech_1997/e97_0681.html

Ellis, N. C. & Bogart, P. S. H. (2007). Speech and language technology in education: The perspective from SLA research and practice. In *Proceedings of the ISCA workshop on Speech and Language Technology in Education (SLaTE 2007)* (pp. 1–8). Farmington, PA, ISCA. Retrieved from https://www.isca-speech.org/archive_open/slate_2007/sle7_001.html

Ellis, R. (1988). The role of practice in classroom language learning. *AILA Review*, *5*, 20–39.

Ellis, R. (2005). Principles of instructed language learning. *System*, *33*(2), 209–224. doi:10.1016/j.system.2004.12.006

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. doi:10.1080/09588221.2011.582845

Engwall, O., Wik, P., Beskow, J. & Granström, B. (2004). Design strategies for a virtual language tutor. In *Proceedings of the International Conference on Spoken Language Processing (InterSpeech 2004)* (pp. 1693–1696). Jeju Island, Korea: ISCA. Retrieved from http://www.isca-speech.org/archive/interspeech_2004/i04_1693.html

Eskenazi, M. (1999). Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language Learning & Technology*, *2*(2), 62–76. doi:10125/25043

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

Eskenazi, M. & Hansma, S. (1998). The Fluency pronunciation trainer. In *Proceedings of the STiLL Workshop*. Retrieved from http://www.cs.cmu.edu/~max/mainpage_files/Esk-Hans-98.pdf

Felix, U. (2005). Analysing recent CALL effectiveness research — Towards a common agenda. *Computer Assisted Language Learning*, *18*(1-2), 1–32. doi:10.1080/09588220500132274

Felshin, S. (1995). The Athena language learning project NLP system: A multilingual system for conversation-based language learning. In V. M. Holland, J. D. Kaplan & M. R. Sams (Eds.), *Intelligent Language Tutors: Theory shaping technology* (pp. 257–272). Mahwah, NJ: Lawrence Erlbaum Associates.

Feuerman, K., Marshall, C., Newman, D. & Rypa, M. (1987). The CALLE project. *CALICO Journal*, *4*(3), 25–34. doi:10.1558/cj.v4i3.25-34

Fruchterman, T. M. J. & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129–1164. doi:10.1002/spe.4380211102

Fryer, L. & Carpenter, R. (2006). Bots as language learning tools. *Language Learning & Technology*, *10*(3), 8–14. doi:10125/44068

Fryer, L. & Nakao, K. (2009). Assessing chatbots for EFL learner use. In A. Stoke (Ed.), *JALT2008 Conference Proceedings* (pp. 849–857). Tokyo: JALT.

Gamper, J. & Knapp, J. (2002). A review of Intelligent CALL systems. *Computer Assisted Language Learning*, *15*(4), 329–342. doi:10.1076/call.15.4.329.8270

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L. & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, *27*(1), 70–105. doi:10.1080/09588221.2012.700315

Griol, D., Baena, I., Molina, J. M. & Sanchis de Miguel, A. (2014). A multimodal conversational agent for personalized language learning. In C. Ramos, P. Novais, C. Nihan & J. M. Corchado Rodríguez (Eds.), *Ambient Intelligence - Software and Applications* (pp. 13–21). doi:10.1007/978-3-319-07596-9_2

Gupta, S., Walker, M. A. & Romano, D. M. (2008). POLLy: A conversational system that uses a shared representation to generate action and social language. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)* (Vol. 2), Hyderabad, India. Retrieved from http://aclweb.org/anthology/I08-2143

Hamburger, H. & Hashim, R. (1992). Foreign language tutoring and learning environment. In M. L. Swartz & M. Yazdani (Eds.), *Intelligent Tutoring Systems for Foreign Language Learning* (80, pp. 201–218). NATO ASI Series. doi:10.1007/978-3-642-77202-3_13

Harless, W. G., Zier, M. A. & Duncan, R. C. (1999). Virtual dialogues with native speakers: The evaluation of an interactive multimedia method. *CALICO Journal*, *16*(3), 313–337. doi:10.1558/cj.v16i3.313-337

Hart, R. S. (1981). The PLATO System and language study. *Studies in Language Learning*, *3*(1), 1–24.

Hassani, K., Nahvi, A. & Ahmadi, A. (2016). Design and implementation of an intelligent virtual environment for improving speaking and listening skills. *Interactive Learning Environments*, *24*(1), 252–271. doi:10.1080/10494820.2013.846265

Hegelheimer, V. & Chapelle, C. A. (2000). Methodological issues in research on learner-computer interactions in CALL. *Language Learning & Technology*, *4*(1), 41–59. Retrieved from http://llt.msu.edu/vol4num1/hegchap/default.html

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

Hjalmarsson, A., Wik, P. & Brusk, J. (2007). Dealing with DEAL: A dialogue system for conversation training. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue* (pp. 132–135). Antwerp: Association for Computational Linguistics. Retrieved from http://www.speech.kth.se/prod/publications/files/3087.pdf

Holden, C. L. & Sykes, J. M. (2011). Leveraging mobile games for place-based language learning. *International Journal of Game-Based Learning*, *1*(2), 1–18. doi:10.4018/ijgbl.2011040101

Holland, V. M., Kaplan, J. D. & Sabol, M. A. (1999). Preliminary tests of language learning in a speech-interactive graphics microworld. *CALICO Journal*, *16*(3), 339–359. doi:10.11139/cj.16.3.339-359

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

Hubbard, P. & Bradin Siskin, C. (2004). Another look at tutorial CALL. *ReCALL*, *16*(02), 448–461. doi:10.1017/S0958344004001326

Jepson, K. (2005). Conversations – and negotiated interaction – in text and voice chat rooms. *Language Learning & Technology*, *9*(3), 79–98. Retrieved from http://www.llt.msu.edu/vol9num3/jepson/default.html

Jia, J. (2009). An AI framework to teach English as a Foreign Language: CSIEC. *AI Magazine*, *30*(2), 59–71. Retrieved from http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/2232

Jia, J. & Chen, W. (2009). The further development of CSIEC project driven by application and evaluation in English education. *British Journal of Educational Technology*, *40*(5), 901–918. doi:10.1111/j.1467-8535.2008.00881.x

Jia, J., Chen, Y., Ding, Z., Bai, Y., Yang, B., Li, M. & Qi, J. (2013). Effects of an intelligent web-based English instruction system on students' academic performance. *Journal of Computer Assisted Learning*, *29*(6), 556–568. doi:10.1111/jcal.12016

Johnson, W. L. & Valente, A. (2009). Tactical Language and Culture Training Systems: Using AI to teach foreign languages and cultures. *AI Magazine*, *30*(2), 72. doi:10.1609/aimag.v30i2.2240

Jokinen, K. & McTear, M. F. (2010). *Spoken dialogue systems* (). Synthesis Lectures on Human Language Technologies. doi:10.2200/s00204ed1v01y200910hlt005

Jurafsky, D. & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Kaplan, J. D., Sabol, M. A., Wisher, R. A. & Seidel, R. J. (1998). The Military Language Tutor (MILT) program: An advanced authoring system. *Computer Assisted Language Learning*, *11*(3), 265–87. doi:10.1076/call.11.3.265.5679

Kim, Y. (2013). Digital peers to help children's text comprehension and perceptions. *Educational Technology & Society*, *16*(4), 59–70.

Kim, N.-Y. (2016). Effects of voice chat on EFL learners' speaking ability according to proficiency levels. *Multimedia-Assisted Language Learning*, *19*(4), 63–88. doi:10.15702/mall.2016.19.4.63

Klüwer, T. (2011). From chatbots to dialog systems. In D. Perez-Marin & I. Pascual-Nieto (Eds.), *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices* (pp. 1–22). doi:10.4018/978-1-60960-617-6.ch001

Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: A conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing*, *16*(2), 163–188. doi:10.1177/026553229901600203

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon Press.

Krüger, A. & Hamilton, S. (1997). RECALL: Individual language tutoring through intelligent error diagnosis. *ReCALL*, *9*(02), 51–58. doi:10.1017/S095834400000478X

Kwon, O.-W., Lee, K., Roh, Y.-H., Huang, J.-X., Choi, S.-K., Kim, Y.-K., ... Lee, Y. (2015). GenieTutor: A computer assisted second-language learning system based on spoken language understanding. In G. G. Lee, H. K. Kim, M. Jeong & J.-H. Kim (Eds.), *Natural Language Dialog Systems and Intelligent Assistants* (pp. 257–262). doi:10.1007/978-3-319-19291-8_26

Lai, C. & Zhao, Y. (2006). Noticing and text-based chat. *Language Learning & Technology*, *10*(3), 102–120. doi:10125/44077

Lech, T. C. & De Smedt, K. (2006). Dreistadt: A language enabled MOO for language learning. In *Proceedings of the ECAI-06 Workshop on Language-enabled Educational Technology*. Retrieved from https://bora.uib.no/handle/1956/1286

Lee, K., Kweon, S.-o., Lee, S., Noh, H. & Lee, G. G. (2014). POSTECH Immersive English Study (POMY): Dialog-based language learning game. *IEICE Transactions on Information and Systems*, *E97-D*(7), 1830–1841. doi:10.1587/transinf.E97.D.1830

Lee, S., Noh, H., Lee, J., Lee, K. & Lee, G. G. (2010). POSTECH approaches for dialog-based English conversation tutoring. In *Proceedings of the Second APSIPA Annual Summit and Conference* (pp. 794–803). doi:10.1.1.407.8645

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. doi:10.1017/s0958344010000273

Lelouche, R. (2008). How the PILÉFACE system, dealing with pragmatics, takes cultural factors into account. In E. G. Blanchard & D. Allard (Eds.), *Proceedings of the Workshop on Culturally Aware Tutoring Systems at ITS 2008 (CATS 2008)* (pp. 47–58). Montreal, Canada. Retrieved from http://hal-lirmm.ccsd.cnrs.fr/lirmm-00351755v2

Levin, L. S. & Evans, D. A. (1995). ALICE-chan: A case study in ICALL theory and practice. In V. M. Holland, J. D. Kaplan & M. R. Sams (Eds.), *Intelligent Language Tutors: Theory shaping technology* (pp. 77–98). Mahwah, NJ: Lawrence Erlbaum.

Lison, P. (2014). *Structured probabilistic modelling for dialogue management* (Doctoral dissertation, University of Oslo, Oslo). Retrieved from http://folk.uio.no/plison/pdfs/thesis/thesis-plison2014.pdf

Lison, P. (2015). A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech & Language*, *34*(1), 232–255. doi:10.1016/j.csl.2015.01.001

Lison, P. & Bibauw, S. (2017). Not all dialogues are created equal: Instance weighting for neural conversational models. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (pp. 384–394). doi:10.18653/v1/w17-5546

Loewen, S. (2011). Focus on form. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (Vol. 2, pp. 576–591). New York, NY: Routledge.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of Second Language Acquisition* (pp. 413–468). San Diego, CA: Academic Press.

Long, M. H. (2015). *Second language acquisition and task-based language teaching* (1st ed.). Malden, MA: Wiley-Blackwell.

Long, M. H. & Robinson, P. (1998). Focus on form: Theory, research and practice. In C. J. Doughty & J. Williams (Eds.), *Focus on form in second language acquisition* (pp. 15–41). The Cambridge Applied Linguistics Series. Cambridge: Cambridge University Press.

Lorenzo, C.-M., Lezcano, L. & Sánchez-Alonso, S. (2013). Language learning in educational virtual worlds - a TAM based assessment. *Journal of Universal Computer Science*, *19*(11), 1615–1637. doi:10.3217/jucs-019-11-1615

Lu, C.-H., Chiou, G.-F., Day, M.-Y., Ong, C.-S. & Hsu, W.-L. (2006). Using instant messaging to provide an intelligent learning environment. In M. Ikeda, K. D. Ashley & T.-W. Chan (Eds.), *Intelligent Tutoring Systems* (pp. 575–583). Lecture Notes in Computer Science. doi:10.1007/11774303_57

MacIntyre, P. D., Dörnyei, Z., Clément, R. & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal*, *82*(4), 545–562. doi:10.1111/j.1540-4781.1998.tb05543.x

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 Applied Linguistics. Oxford: Oxford University Press.

Massaro, D. W., Liu, Y., Chen, T. H. & Perfetti, C. (2006). A multilingual embodied conversational agent for tutoring speech and language learning. In *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP, September, Pittsburgh, PA)* (pp. 825–828). Bonn: Universität Bonn. Retrieved from http://www.pitt.edu/~perfetti/PDF/Multilingual%20embodied%20conversational%20agent-%20Massaro%20et%20al..pdf

Molla, S. R., Sanders, A. F. & Sanders, R. H. (1988). Artificial intelligence in a German adventure game: Spion in PROLOG. *CALICO Journal*, *6*(1), 9–23. doi:10.11139/cj.6.1.9-23

Morton, H., Davidson, N. & Jack, M. A. (2008). Evaluation of a speech interactive CALL system. In F. Zhang & B. Barber (Eds.), *Handbook of Research on Computer-Enhanced Language Acquisition and Learning* (pp. 219–239). doi:10.4018/978-1-59904-895-6.ch013

Morton, H., Gunson, N. & Jack, M. A. (2012). Interactive language learning through speech-enabled virtual scenarios. *Advances in Human-Computer Interaction*, *2012*, e389523. doi:10.1155/2012/389523

Morton, H. & Jack, M. A. (2005). Scenario-based spoken interaction with virtual agents. *Computer Assisted Language Learning*, *18*(3), 171–191. doi:10.1080/09588220500173344

Muranoi, H. (2007). Output practice in the L2 classroom. In R. M. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 51–84). Cambridge; New York: Cambridge University Press.

Norris, J. M. & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language learning*, *50*(3), 417–528. doi:10.1111/0023-8333.00136

Norris, J. M., Ross, S. J. & Schoonen, R. (Eds.). (2015). *Improving and extending quantitative reasoning in second language research*. Malden, MA: Wiley.

Ortega, L. (2007). Meaningful L2 practice in foreign language classrooms: A cognitive-interactionist SLA perspective. In R. M. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 180–207). Cambridge; New York: Cambridge University Press.

Otte, E. & Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information sciences. *Journal of Information Science*, *28*(6), 441–453. doi:10.1177/016555150202800601

Petersen, K. A. (2010). *Implicit corrective feedback in computer-guided interaction: Does mode matter?* (Doctoral dissertation, Georgetown University, Washington, DC). doi:10822/553155

Pica, T. (2013). From input, output and comprehension to negotiation, evidence, and attention: An overview of theory and research on learner interaction and SLA. In M. d. P. García Mayo, M. J. Gutierrez Mangado & M. Martínez Adrián (Eds.), *Contemporary Approaches to Second Language Acquisition* (pp. 49–70). AILA Applied Linguistics Series. doi:10.1017/CBO9781139524735.006

Piwek, P. (2017). Dialogue with computers: Dialogue games in action. In J. Mildorf & B. Thomas (Eds.), *Dialogue across media* (28, pp. 179–202).

Dialogue Studies. Amsterdam: John Benjamins. Retrieved from https://dx.doi.org/10.1075/ds.28.10piw

Plonsky, L. & Gass, S. M. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*(2), 325–366. doi:10.1111/j.1467-9922.2011.00640.x

Plonsky, L. & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. doi:10.1111/lang.12079

Price, C., Bunt, A. & McCalla, G. (1999). L2tutor: A mixed-initiative dialogue system for improving fluency. *Computer Assisted Language Learning*, *12*(2), 83–112.

Quixal, M. & Meurers, D. (2016). How can writing tasks be characterized in a way serving pedagogical goals and automatic analysis needs? *CALICO Journal*, *33*(1), 19–48. doi:10.1558/cj.v33i1.26543

Raux, A. & Eskenazi, M. (2004). Using task-oriented spoken dialogue systems for language learning: Potential, practical applications and challenges. In *Proceedings of INSTIL 2004*, Venice, Italy: ISCA. Retrieved from http://www.isca-speech.org/archive_open/icall2004/iic4_035.html

Rayner, E., Tsourakis, N., Baur, C., Bouillon, P. & Gerlach, J. (2012). CALL-SLT: A Spoken CALL System based on grammar and speech recognition. *Linguistic Issues in Language Technology (LiLT)*, *10*(2). Retrieved from http://csli-lilt.stanford.edu/ojs/index.php/LiLT/article/view/15

Read, T. (2014). The architectonics of language MOOCs. In E. Martín-Monje & E. Bárcena (Eds.), *Language MOOCs: Providing learning, transcending boundaries* (pp. 91–105). doi:10.2478/9783110420067.6

Robinson, P. (Ed.). (2011). *Second language task complexity: Researching the cognition hypothesis of language learning and performance*. Amsterdam: John Benjamins.

Rypa, M. (1996). VILTS: The voice interactive language training system. In F. Borchardt & E. M. Johnson (Eds.), *Proceedings of the CALICO 1996 Annual Symposium*, Durham, NC: CALICO. Retrieved from https://www.sri.com/sites/default/files/publications/vilts-the_voice_interactive_language_training_system.pdf

Rypa, M. & Price, P. (1999). VILTS: A tale of two technologies. *CALICO Journal*, *16*(3), 385–404. doi:10.1558/cj.v16i3.385-404

Saerbeck, M., Schut, T., Bartneck, C. & Janse, M. D. (2010). Expressive robots in education: Varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1613–1622). doi:10.1145/1753326.1753567

Sanders, R. H. & Sanders, A. F. (1995). History of an AI spy game: Spion. In V. M. Holland, J. D. Kaplan & M. R. Sams (Eds.), *Intelligent Language Tutors: Theory shaping technology* (pp. 141–152). Mahwah, NJ: Lawrence Erlbaum Associates.

Schoelles, M. & Hamburger, H. (1997). The NLP role in animated conversation for CALL. In *Proceedings of the Fifth Applied Natural Language Processing Conference (ANLP 1997)* (pp. 127–134). Washington, USA: ACL. Retrieved from http://aclweb.org/anthology-new/A/A97/A97-1019.pdf

Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P. & Zue, V. W. (1998). GALAXY-II: A reference architecture for conversational system development. In *Proceedings of ICSLP* (pp. 931–934). Sydney, Australia: ISCA. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.40.3907

Seneff, S., Wang, C. & Chao, C.-y. (2007). Spoken dialogue systems for language learning. In *Proceedings of Human Language Technologies: The Annual*

Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (pp. 13–14). NAACL-Demonstrations '07. Stroudsburg, PA, USA: ACL. Retrieved from https://dl.acm.org/citation.cfm?id=1614171

Seneff, S., Wang, C., Peabody, M. A. & Zue, V. W. (2004). Second language acquisition through human computer dialogue. In *Proceedings of the International Symposium on Chinese Spoken Language Processing 2004* (pp. 341–344). ISCSLP 2004. doi:10.1109/CHINSL.2004.1409656

Sha, G. (2009). AI-based chatterbots and spoken English teaching: A critical analysis. *Computer Assisted Language Learning*, *22*(3), 269–281. doi:10.1080/09588220902920284

Singh, S., Litman, D., Kearns, M. & Walker, M. A. (2002). Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. *Journal of Artificial Intelligence Research*, *16*, 105–133. doi:10.1613/jair.859. arXiv: 1106.0676

Stewart, I. A. D. & File, P. (2007). Let's Chat: A conversational dialogue system for second language practice. *Computer Assisted Language Learning*, *20*(2), 97–116. doi:10.1080/09588220701331386

Stokes, P. D. (2005). *Creativity from constraints: The psychology of breakthrough*. New York: Springer.

Su, P.-H., Wang, Y.-B., Yu, T.-h. & Lee, L.-S. (2013). A dialogue game framework with personalized training using reinforcement learning for computer-assisted language learning. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)* (pp. 8213–8217). doi:10.1109/icassp.2013.6639266

Su, P.-H., Wu, C.-h. & Lee, L.-s. (2015). A recursive dialogue game for personalized computer-aided pronunciation training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(1), 127–141. doi:10.1109/TASLP.2014.2375572

Surface, E. A., Dierdorff, E. C. & Watson, A. M. (2007). *Special operations language training software measurement of effectiveness study: Tactical Iraqi study final report* (Technical report No. 2007010602). SWA Consulting. Retrieved from https://rdl.train.army.mil/catalog/view/100.ATSC/56F12B52-936B-4B85-A11D-44D3F590E2DF-1274550929321/BCKStest/ussocom_tactical_iraqi_study_final_report_swa_20070501.pdf

Suzuki, N., Nose, T., Hiroi, Y. & Ito, A. (2014). Controlling switching pause using an AR agent for interactive CALL system. In C. Stephanidis (Ed.), *HCI International 2014 - Posters' Extended Abstracts - Proceedings, Part II* (pp. 588–593). Communications in Computer and Information Science. doi:10.1007/978-3-319-07854-0_102

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471–483). Mahwah, NJ: Lawrence Erlbaum.

Sydorenko, T. (2015). The use of computer-delivered structured tasks in pragmatic instruction: An exploratory study. *Intercultural Pragmatics*, *12*(3), 333–362. doi:10.1515/ip-2015-0017

Sydorenko, T., Smits, T. F. H., Evanini, K. & Ramanarayanan, V. (2018). Simulated speaking environments for language learning: Insights from three cases. *Computer Assisted Language Learning*, *0*(0), 1–32. doi:10.1080/09588221.2018.1466811

Tegos, S., Demetriadis, S. & Karakostas, A. (2013). MentorChat: A teacher-configurable conversational agent that promotes students' productive talk. In D. Hernández-Leo, T. Ley, R. Klamma & A. Harrer (Eds.), *Scaling*

up Learning for Sustained Impact (8095, pp. 581–584). Lecture Notes in Computer Science. doi:10.1007/978-3-642-40814-4_63

Tin, T. B. (2012). Freedom, constraints and creativity in language learning tasks: New task features. *Innovation in Language Learning and Teaching*, *6*(2), 177–186. doi:10.1080/17501229.2011.628024

Underwood, J. H. (1982). Simulated conversation as a CAI strategy. *Foreign Language Annals*, *15*(3), 209–212. doi:10.1111/j.1944-9720.1982.tb00248.x

Underwood, J. H. (1984). *Linguistics, computers and the language teacher: A communicative approach*. Rowley, MA: Newbury House.

Vaassen, F., Wauters, J., Van Broeckhoven, F., Van Overveldt, M., Eneman, K. & Daelemans, W. (2012). deLearyous: Training interpersonal communication skills using unconstrained text input. In P. Felicia (Ed.), *Proceedings of ECGBL 2012, The 6th European Conference on Games Based Learning*, Cork, Ireland: Academic Publishing International.

Venkatesh, V. & Davis, F. D. (2000). A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science*, *46*(2), 186–204. doi:10.1287/mnsc.46.2.186.11926

Vinyals, O. & Le, Q. (2015). A neural conversational model. In *Proceedings of the ICML Deep Learning Workshop 2015*, Lille, France: IMLS. arXiv: 1506.05869. Retrieved from https://sites.google.com/site/deeplearning2015/36.pdf?attredirects=0

Vlugter, P., Knott, A., McDonald, J. & Hall, C. (2009). Dialogue-based CALL: A case study on teaching pronouns. *Computer Assisted Language Learning*, *22*(2), 115–131. doi:10.1080/09588220902778260

Wachowicz, K. A. & Scott, B. (1999). Software that listens: It's not a question of whether, it's a question of how. *CALICO Journal*, *16*(3), 253–276. doi:10.1558/cj.v16i3.253-276

Walker, N. R., Trofimovich, P., Cedergren, H. & Gatbonton, E. (2011). Using ASR technology in language training for specific purposes: A perspective from Quebec, Canada. *CALICO Journal*, *28*(3), 721–743. doi:10.11139/cj.28.3.721-743

Wallace, R. S. (2003). *The elements of AIML style*. ALICE AI Foundation. Retrieved from http://www.alicebot.org/style.pdf

Wang, C. & Seneff, S. (2007). A spoken translation game for second language learning. In *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work* (pp. 315–322). Amsterdam: IOS Press. Retrieved from http://dl.acm.org/citation.cfm?id=1563601.1563653

Wang, N. & Johnson, W. L. (2008). The politeness effect in an intelligent foreign language tutoring system. In B. P. Woolf, E. Aïmeur, R. Nkambou & S. Lajoie (Eds.), *Proceedings of ITS 2008: Intelligent Tutoring Systems* (5091, pp. 270–280). Lecture Notes in Computer Science. doi:10.1007/978-3-540-69132-7_31

Wang, Y. F., Petrina, S. & Feng, F. (2017). VILLAGE–Virtual immersive language learning and gaming environment: Immersion and presence. *British Journal of Educational Technology*, *48*(2), 431–450. doi:10.1111/bjet.12388

Weizenbaum, J. (1966). ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

Williams, L. & van Compernolle, R. A. (2009). The chatbot as a peer/tool for learners of French. In G. Lord & L. Lomicka (Eds.), *The next genera-*

tion: Social networking and online collaboration in foreign language learning (pp. 145–172). San Marcos, TX: CALICO.

Wilske, S. (2015). *Form and meaning in dialog-based computer-assisted language learning* (Doctoral dissertation, Universität des Saarlandes, Saarbrücken). Retrieved from http://scidok.sulb.uni-saarland.de/volltexte/2015/6251/

Wilske, S. & Wolska, M. (2011). Meaning versus form in computer-assisted task-based language learning: A case study on the German dative. *Journal for Language Technology and Computational Linguistics*, *26*(1), 23–37. Retrieved from http://www.jlcl.org/2011_Heft1/H2011-1.pdf#page=28

Young, R. (1988). Computer-assisted language learning conversations: Negotiating an outcome. *CALICO Journal*, *5*(3), 65–83. Retrieved from http://www.equinoxpub.com/journals/index.php/CALICO/article/viewArticle/23558

Young, S. (2000). Probabilistic methods in spoken–dialogue systems. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *358*(1769), 1389–1402. Retrieved from http://rsta.royalsocietypublishing.org/content/358/1769/1389.short

Young, S., Gašić, M., Thomson, B. & Williams, J. D. (2013). POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, *101*(5), 1160–1179. doi:10.1109/JPROC.2012.2225812

Ziegler, N. (2016). Synchronous computer-mediated communication and interaction: A meta-analysis. *Studies in Second Language Acquisition*, *38*(3), 553–586. doi:10.1017/S027226311500025X

SUPPLEMENTARY ANNEXES

Annex I: Complete list of publications collected for this study (343 publications).

Annex II: List of publications on dialogue-based CALL, with all coded variables (207 publications).

Annex III: List of dialogue-based CALL systems, with all coded variables (96 systems).

Annex IV: Coding scheme for publications.

Annex V: Coding scheme for systems.