http://www.diva-portal.org

Postprint

This is the accepted version of a paper published in . This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

Engwall, O., Águas Lopes, J D., Cumbal, R., Berndtson, G., Lindström, R. et al. [Year unknown!]
Learner and teacher perspectives on robot-led L2 conversation practice
[Journal name unknown!]

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-305048

Learner and teacher perspectives on robot-led L2 conversation practice

Olov Engwall
KTH Royal Institute of Technology, Sweden (engwall@kth.se)

José Lopes
Heriot-Watt University, Edinburgh, United Kingdom (jd.lopes@hw.ac.uk)

Ronald Cumbal
KTH Royal Institute of Technology, Sweden (ronaldcg@kth.se)

Gustav Berndtson
KTH Royal Institute of Technology, Sweden (berndtso@kth.se)

Ruben Lindström
KTH Royal Institute of Technology, Sweden (rubenli@kth.se)

Patrik Ekman
KTH Royal Institute of Technology, Sweden (paekm@kth.se)

Eric Hartmanis
KTH Royal Institute of Technology, Sweden (erichar@kth.se)

Emelie Jin
KTH Royal Institute of Technology, Sweden (ejin@kth.se)

Ella Johnston
KTH Royal Institute of Technology, Sweden (ellajo@kth.se)

Gara Tahir
KTH Royal Institute of Technology, Sweden (garat@kth.se)

Michael Mekonnen
KTH Royal Institute of Technology, Sweden (micmek@kth.se)

Abstract

This article focuses on designing and evaluating conversation practice in a second language (L2) with a robot that employs human spoken and non-verbal interaction strategies. Based on an analysis of previous work and semi-structured interviews with L2 learners and teachers, recommendations for robot-led conversation practice for adult learners at intermediate level are first defined, focused on language learning, on the social context, on the conversational structure and on verbal and visual aspects of the robot moderation. Guided by these recommendations, an experiment is set up, in which 12 pairs of L2 learners of Swedish interact with a robot in short social conversations. These robot-learner interactions are evaluated through post-session interviews with the learners, teachers' ratings of the robot's behaviour and analyses of the video-recorded conversations, resulting in a set of guidelines for robot-led conversation practice, in particular: 1) Societal and personal topics increase the practice's meaningfulness for learners. 2) Strategies and methods for providing corrective feedback during conversation practice need to be explored further. 3) Learners should be encouraged to support each other if the robot has difficulties adapting to their linguistic level. 4) The robot should establish a social relationship, by contributing with its own story, remembering the participants' input, and making use of non-verbal communication signals. 5) Improvements are required regarding naturalness and intelligibility of text-to-speech synthesis, in particular its speed, if it is to be used for conversations with L2 learners.

Keywords: educational robots; conversation practice; L2 speaking; multi-party interaction

1. Introduction

Automating spoken conversation practice in computer-assisted language learning (CALL) of a second language (L2) learning is a challenging task, both pedagogically and technologically. The primary pedagogical challenge is to create a practice that is both stimulating – motivating learners to interact with a chatbot, a computer-animated character or a robot – and realistic – allowing the practice outcome to be transferred to real-world situations. The main technological challenge is to autonomously interpret conversational non-native learner utterances since these may deviate semantically and phonetically from the standard. Previous CALL systems for conversation practice have often restricted the conversations to specific role-playing scenarios with computer-animated agents in virtual environments, in which the learner input is either provided by the system for the learner to read or is more easily predicted by a narrow task-based context. This allows the system to correctly interpret the learners' input, and possibly provide corrective feedback using comparisons to the expected input. Examples for adult L2 learners include the 3D virtual reality environment and task scenarios such as "at the restaurant" created by Morton & Jack (2005) for beginner learners of French, Italian, Japanese and English, and the role-play dialogues with a virtual shopkeeper at a flea market that Wik & Hjalmarsson (2009) used to investigate conversation practice of Swedish. Johnson & Valente (2008) and Hautopp & Hanghøj (2014) created similar 3D computer-animated scenarios in which the learners, respectively US army recruits and immigrants to Denmark, could interact in freer spoken conversation with virtual version of their target community to practice the L2 (Arabic or Danish) and cultural norms (in Iraq and Denmark, respectively). Such conversation practice is intended to be transferable to real-life situations. However, the screen-based virtual environments lack in realism regarding the physical rapport of situated, often multi-party, interaction that is common in real-life conversations. Educational robots may therefore improve conversation practice by strengthening social aspects, regarding e.g., turn-taking, non-verbal signals, and mutual peer learner support in multi-learner settings.

The extensive reviews by van den Berghe *et al.* (2018) and Randall (2019) of, respectively, 33 and 79 studies on using social robots to practice vocabulary, grammar, reading, speaking and sign language, show that robots may contribute to effective learning, e.g., by increasing learner motivation, time on practice or self-confidence, or by introducing alternative methods for learning. However, as pointed out by Neri *et al.* (2002), there is a risk and tendency that CALL development is technology-driven rather than pedagogically founded, i.e., implementation is based on what is technologically possible, instead of what is most beneficial for learning. This study therefore uses data triangulation of several different learning-focused data sources to answer the research question: *What recommendations for robot-led social L2 conversation practice for adult learners at intermediate level do learners, teachers and researchers give based on experience and a user test?*

To do so, a literature review of previous work first identifies important aspects to consider for robot-led L2 conversations (Section 2). These aspects guide a questionnaire and interviews with L2 teachers and learners (Section 3) to determine pedagogical requirements and expectations on robot-led conversation practice similar to that at language cafés. These are informal gatherings in which L2 speakers meet L1 speakers to practice listening skills and fluency through social small talk, e.g., about home countries, languages, society, and hobbies. The identified requirements are in turn considered in the implementation of a framework for robot-led L2 conversations (Section 4). A user experiment is conducted (Section 5), after which the interaction is evaluated by interviews with learners (Section 6.1), ratings of the robot's behaviour by teachers (Section 6.2) and interaction analysis of the video-recorded conversations to identify current shortcomings in the robot's interaction (Section 6.3). Evaluation by both students and teachers is used since it has been shown (e.g., Kirschner & van Merriënboer, 2013) that students are not always aware of what is most beneficial for their learning. Together, these evaluations provide a more complete pedagogical perspective of strengths, weaknesses and required future work. Many aspects are relevant for CALL-based conversation practice in general (Section 7), regardless of whether the intelligent tutor is a spoken chatbot, an animated conversational character or a robot.

2. Literature review: Previous work on L2 conversation practice with robots

Almost all previous studies of conversation practice with robots are with children and our review therefore includes studies with children, even though the present work is with adults. A summary of differences related to age for how different robots, robot roles and teaching strategies may be combined is presented in Engwall & Lopes (2020). The review concluded that adult learners require real-life practice to increase their intrinsic motivation. This can be achieved with teaching strategies such as communicative language teaching, collaborative language learning, or task-based language teaching.

To our knowledge, the only other previous robot-led conversation practice study with adults is Khalifa *et al.* (2017), who used a three-party setting with one Nao robot asking another robot and a human student similar questions and studied to what extent the student learned new English expressions from the robot. Engwall *et al.* (2020) presented the background for and implementation of robot-led L2 conversation practice for adult L2 learners of Swedish, inspired by language café interactions. From that study, four different areas to investigate further are identified. Firstly, that language café interactions provide a **social context** that allow for realistic practice. Secondly, that the **conversational interaction** of language cafés is quite specific and mix different types of interaction. An important reason for this is to address different aspects of **language learning**, such as adapting to the differences in linguistic level of heterogeneous learner groups and distributing feedback in a constructive manner during conversation practice. User tests with robot-led conversation practice showed that aspects that are specific for **robot moderation** – including text-to-speech synthesis (TTS), automatic speech recognition (ASR) and visual appearance – influenced the learners' interaction with the robot and their perception of how personal and human-like it was.

These four dimensions will be used throughout this article and previous findings in relation to each are presented below.

Language learning: Motivation is one of the most important aspects for learning and Shin & Shin (2015) showed that participation and satisfaction were significantly higher for a Korean middle school class working with a robot than for another class using traditional CALL. Learner engagement may, however, be lost if the practice is too hard or too easy, as shown by Baxter et al. (2017) using robot tutors practising math and history material with preschool children. van den Berghe et al. (2018) and Randall (2019) further found that results in previous studies are mixed regarding how robots' social behaviour affects language learning. One reason for this may be a mismatch between the robot's social capabilities and the learners' expectations, which may cause social interaction to interfere with learning. Randall (2019) therefore argued that the robot role should be that of peer, not teacher, since limitations in analysis and feedback capabilities make it difficult for the robot to meet expectations on how teachers should interact with learners.

Social context: Belpaeme *et al.* (2018) highlighted the importance of familiar and engaging overall themes in their analysis of requirements for using robots in L2 tutoring of pre-school children. The social context further includes the relation to peers: Wang et al. (2013), who used robot puppets to practice greetings and self-presentations in English conversations and singing with Taiwanese fifth graders, found that collaborative interaction between peers with the robot (*co-discovery*) was more effective for learning than sequential individual interaction, in which each learner was instructed by the preceding (*peer tutoring*). The post-test showed that the *co-discovery* group improved more than the *peer tutoring* group. In addition, both robot groups had a more positive attitude, less anxiety, and higher motivation and confidence in learning English than a group learning without a robot. Westlund *et al.* (2016) similarly observed that learners were less anxious about interacting with the robot if they had familiarised themselves with it together with peers.

Conversational interaction: Engwall *et al.* (2020) found that the 105 language café moderators who answered a questionnaire responded that they mainly *interview* one or several participants about a topic; *answer* participants' questions; *narrate,* e.g., about the country; *facilitate* by initiating interaction between participants; or *assist* in their conversation. The actual strategy use was then analysed in 14 short human-led conversations with one moderator and two L2 learners. Based on the observations, four robot interaction strategies (*Interviewer, Narrator, Facilitator* and *Interlocutor*) were implemented for a user test aimed at investigating how, e.g., the learners' linguistic level, gender, cultural origin and familiarity with the peer influence preferences for how the robot should structure the conversation. In the present study, a new user test is conducted to gain additional insights on pedagogical and technological challenges.

Robot moderation: Studies by Lee *et al.* (2011), In & Han (2015) and Hong *et al.* (2016) have shown that TTS may pose problems for robot-assisted language learning. Lee *et al.* (2011) found that TTS was not of sufficient quality for practicing listening skills in role-play scenarios between Korean middle-school students and a robot. In the post-tests, learners had improved their speaking skills, but not their oral comprehension. Hong *et al.* (2016) on the contrary found that fifth-grade Taiwanese students interacting in various teaching activities with a humanoid robot had improved their L2 listening skills more than the control group, but not their L2 speaking skills. In & Han (2015) let a robot teach declarative *vs*. interrogative intonation to 10-year-old Korean children but concluded that the paralinguistic features of the TTS were not distinct enough to teach prosody.

Starting from these insights from previous work, an exploratory pre-study was conducted.

3. Pre-study: requirements for robot-led L2 conversation practice

The pre-study consisted of two parts: one questionnaire including free-text questions sent out to a larger group of L2 Swedish teachers and two sets of semi-structured interviews with teachers, researchers, and learners. This combination was selected to gather both broad (questionnaire) and deeper (interviews) perspectives. The questionnaire, answered by 27 L2 Swedish teachers and language café moderators, focused on the use of robots in conversation practice. The first set of interviews was with two L2 teachers organising language cafés and two researchers, respectively specialised in language and social interaction, specifically how language café moderators and participants co-create the learning environment (Ali Reza Majlesi, Stockholm University), and Swedish as a second language, specifically in pedagogy for spoken practice (Elisabeth Zetterholm, Stockholm University). The relevant questions are summarised in the left part of Table 1. The second set was with 13 language café participants (7 female, 6 male, median age 25 years, representing a variety of L1-L2 pairs {Mauritian, Somali, Kyrgyz}–Swedish, {Spanish, English}–German, {Norwegian, French}–French, Swedish–Chinese, with both beginner to intermediate level learners and native speaker hosts). It focused on the questions shown in the right part of Table 1. The answers are listed below, with indications if they were mentioned by teachers (T), researchers (R) or language café participants (L) and the number of times. The data was analysed inductively, with all answers being summarised and categorised, before being grouped theme-wise post-analysis.

Language learning: Finding an appropriate level is a challenge in language café settings since the learners' levels differ (1L), making it difficult to create practice adequate for all (1L). This is often mitigated by encouraging more proficient peers to assist lower levels learners (8L). Corrective feedback is expected (1R, 5L), at least when it benefits the whole group (1L) or if the moderator uses recasts to reformulate erroneous learner utterances rather than explicit corrections (1L). However, some learners did not want to give or receive feedback in conversation settings (2L) and it was estimated that human moderators correct only about 20% of the errors (1R+1T). Regarding using robots as moderators, opinions were divided, between those who saw no advantages (6T+2L) and those who were positive (22T+2R+4L), because robots could be motivating to work with (3T+4L), that they could increase training time and availability (10T+3L), that they would be objective when correcting learner errors (3T+3L) or would be less intimidating to practice with (3T+2L). Robots should adapt to the learners' linguistic level (2T+1L), but still challenge them (3T+2L) and assist them with linguistic problems (1L). Their spoken utterances should be linguistically correct (1L) and clearly produced (6T+2L), as a pronunciation target and to be easily understandable. They should make sure that everyone understands or else reformulate utterances (4T+1L).

Social context: The focus should be firstly on personal and familiar topics (11L), to promote the social atmosphere and to allow learners to concentrate on utterance formulation (1L), and secondly on societal content, since learning about the society the learners live in is as important as language practice (1T+1R+7L). Commonly occurring and appreciated topics include culture and cultural differences (7L), everyday life (6L), self-presentations and own experiences (6L), the news (3L) and languages (2L). It was further stressed that diversity in topics is a key feature (2L). Social interaction with peers is the most important aspect of language cafés (2T+3L) and the lack of social interaction was deemed to be the largest disadvantage of using robot moderators (14T+8L). Robots do not allow learners to meet native speakers who assist them with matters outside the language café, such as paperwork and homework (1R+1L), or who engage in social activities, e.g., baking together (2L). Since respondents stated that robots could never replace human moderators (2T+2L) or that it had to be better than a human moderator (1L), it is important to clearly present robot-led conversations as a complement (e.g., in case of low availability of L1 speakers or to gain confidence and L2 proficiency). Moreover, whereas some learners stated that they would prefer to use the robot individually as a personal coach (2L), more respondents stressed that human-human interaction is required to establish a social atmosphere (12T+9L). This suggests that a multiparty setting, with one robot and several learners should be used and that learners should be encouraged to also interact with each other. The multiparty setting could further facilitate familiarisation with the robot, as some learners stated that they would need getting used to the robot before engaging with it in conversation practice (3L), while some would be comfortable directly (2L).

Other mentioned social disadvantages of using robots were that they would not have cognition and feelings (3T+4L), that they would be impersonal (4T+2L) and lack cultural knowledge (1L). While cognition may be hard to achieve in robots, it is hence important that they should establish a personal relationship by remembering participants (1L), their names and previous input (2L) and by providing own stories and opinions.

Conversational interaction: Language café interactions are different from the L2 classroom (2T+2R+2L) since it depends on the learners' level (2T+2L), if they are first time attendees or returning (2L) and on the moderator (1L). In general, standard, basic questions are used the first time, but at subsequent meetings, interaction develops into freer and deeper discussions (2L), in which learners get the opportunity to talk as much as possible (2L). This development was stressed as important, since questions-and-answers practice becomes too constrained (2T+3L).

The L1 speakers' role at language cafés varies substantially, from no moderator responsibilities (4L), over being more active to initiate discussions when needed (9L) to a formal role to lead and support conversations (2L). A good moderator should be helpful and provide linguistic support (4L), be patient (3T+1R+3L), present topics and ask questions around it (1T+3L) but know when to leave the initiative to learners (4L) and distribute the word to involve passive participants (1L).

Robot moderation: Robot moderators should be human-like (4T+1L), but after having been shown a video of robot-human interaction, learners were divided regarding whether the synthetic speech was adequate (2L) or monotonous (3L) and whether the facial expressions were good (2L) or unnerving (4L), concerning the appearance of the eyes (2L) and head movements (1L). Fears were that the robot would not understand learners (2T+4L), due to lack of contextual knowledge (1L) or learners' deviating pronunciation (2T+1R).

When summarising the expectations for robot moderators, some appear currently beyond state-of-the-art (constituting a clearer pronunciation target than humans, being able to correct all learner errors and provide linguistic support for learner difficulties) but others are within reach. In fact, after seeing the video many respondents stated that the robot met (11T+1L) or exceeded (1T+6L) their expectations.

Table 1. Thematically reordered important questions for the pre-study interviews.

Language café organisers & researchers	**Language café participants**
Language learning	
For what type of learning is language cafés well, or not, suited?	What is important for a successful language café conversation? Why?
Social context	
What is the contribution of language cafés? Describe typical language café participants. Are some returning?	How often do you visit language cafés? Describe your latest language café visit.
Conversational interaction	
Are conversations free or moderated? What language café topics are common?	How should moderators behave? What moderator qualities are important? How much do you talk with the moderator compared to with peers? What topics are interesting to discuss? Why?
Robot moderation	
When could a robot moderator be appropriate? What would be most important in robot-led conversations? What advantages and difficulties do you see of having robot moderators?	How would you feel about robot moderators? How should a robot moderator behave? What (dis)advantages do you see of having robot moderators? [Picture shown] What are your first impressions of and expectations on this robot? [Video shown] Did it meet your expectations? What could be improved?

4. Implementation in robot-led conversation practice

Based on the findings of Sections 2-3, we summarise implementation choices for the present study as:

Language learning: To improve adaptation to learners' level, the robot is controlled by a semi-automated wizard-of-Oz set-up (c.f. Section 5), allowing the wizard to monitor learner problems and change topic or addressee when problems occur. The choice of using a wizard is based on the current limitations of ASR for L2 conversational speech.

All robot utterances are pre-generated, which means that the robot cannot provide linguistic support or reformulate utterances that were not understood.

Learner errors are not corrected, both for pedagogical reasons of not interrupting the conversation to promote fluency and self-confidence in the L2, and technological challenges of automatically generating corrective feedback in free conversations. Encouraging positive feedback is instead provided.

Social context: To ensure that the practice is familiar and transferable to human-human interaction, conversations are focused on typical everyday social topics, such as presenting oneself (background, occupation, hobbies etc), talking about Sweden and Swedish and comparing to the learners' home countries and languages. To establish personal relationships, the robot provides (factual or authored) information about itself and refers to learners, their home countries, and native languages by name.

A set-up with learner pairs is used, to achieve pedagogical advantages (that the peers may support each other linguistically) as well as technological (that peer interaction may save the practice in situations if technology fails) and social (that peer interaction may improve the conversation and reduce learner anxiety of talking with a robot). The robot encourages peer interaction by asking questions that both learners need to provide input to (e.g., "*Do you live close to each other?*"), requesting peer comments ("*Do you also like that movie?*"), enticing collaboration ("*Do you two know which the four largest cities in Sweden are?*") or encouraging peer assistance ("*Could you help explaining?*").

Conversational interaction: The robot initiates and drives the conversation in the three-party interaction, by asking questions directed to one individual learner or to both and providing its own views. Limited interaction adjustments are made depending on the participants' activity. Compared to Engwall *et al.* (2020) the interaction contains more individual follow-up questions to each learner and robot answers for counter-questions by learners, since individual interviewing was the most appreciated by learners and learner questions were frequent.

Robot moderation: Default TTS synthesis, with a Swedish voice from CereProc, is used for the robot's vocal utterances, lip movements and facial expressions on the computer-animated face. This means that no manual adaptations are made of the robot's voice (speed, intonation), facial expressions (emotional, word emphasis), or eye gaze patterns (other than turning towards the addressed learner).

The above implementation choices were tested in the experiments described in Section 5.

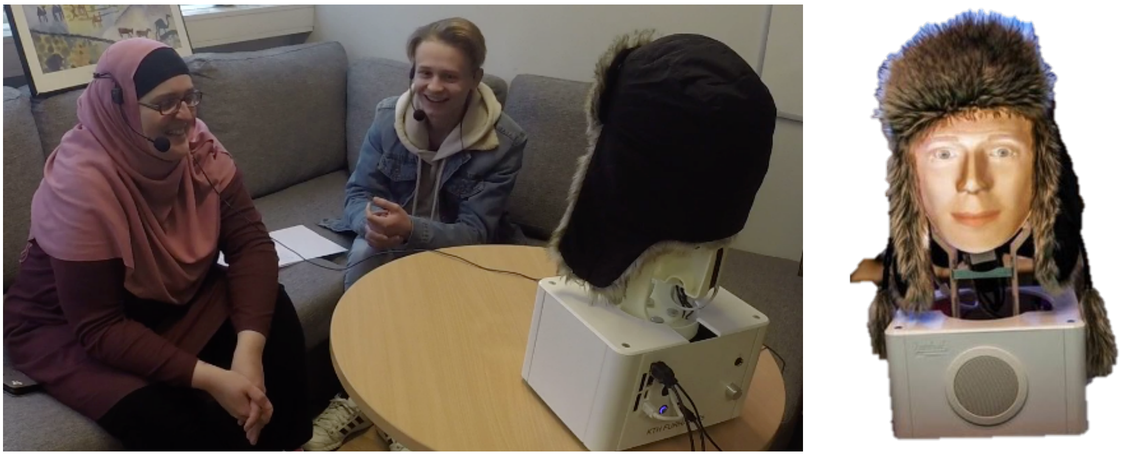


Figure 1. Left: Setup of the robot-led conversation. Right: Furhat robot, with computer-animated face back-projected on 3D face mask.

5. Experimental set-up

We conducted 12 robot-led conversations of approximately 10 minutes each with 12 pairs of adult L2 learners (17 women, 7 men) attending a Swedish for Immigrants course at level B2–C1. As is normally the case in such courses, the learners had a large variety in L1 background (3 Arabic, 2 Chinese, 2 Polish, Russian, Ukrainian, Belarussian, Azerbaijan, Armenian, Serbian, Bosnian, German, Greek, Italian, Indonesian, Tamil, Hindi, Dari, Urdu, Persian), in age (24–60 years) and in how long they had been studying Swedish (between 1 month and 4 years). The robot Furhat (Al Moubayed *et al.*, 2012) was placed on a table facing the learners, who were seated on a sofa (c.f. Figure 1). An informed consent form was used to describe the aim of the study and the use of collected data, but learners were not otherwise instructed regarding how to interact with the robot. To avoid that later learner pairs' expectations were influenced, the wizard-of-Oz set-up was not disclosed until all conversations had been conducted. The wizard controlled the dialogue flow using predefined utterances that were automatically updated at each robot turn, based on the previous robot utterance, so that the wizard could choose the next robot utterance from general statements (*Yes/No/I do not really know*), backchannels and feedback (*Mm/Mhm/Interesting*) and up to nine utterances with follow-up questions, topic transitions and robot answers to counter-questions. The wizard was seated in the same room, but hidden behind an office divider, so that he could hear the learners' input and based on this choose the most appropriate next robot utterance using shortcut keys. To allow for post-session analysis, the conversations were recorded using a GoPro video camera capturing the scene (as shown in Figure 1) and one headset microphone per participant.

6. Evaluation

The 12 conversations were transcribed and analysed in three ways: a) through semi-structured interviews with the 24 learners (Section 6.1), b) via teacher assessment of the robot's interaction (Section 6.2) and c) by an interaction analysis of the video recordings (Section 6.3).

6.1 Post-interaction interviews with learners

After each conversation, learner pairs were interviewed during 10-20 minutes about the session, in Swedish and/or English, depending on their preferences. Table 2 lists questions of the semi-structured interview reordered thematically post-analysis for this article together with common answers. An inductive data analysis was employed, i.e., all interviews were recorded and thereafter analysed as a set, to identify common responses and count their frequency. Some responses emanate directly from one question (e.g., "*Were there any unnatural pauses?*"), whereas others were offered to more general questions, as indicated by Table 2. To quantify the learner responses, the number of learners providing positive or negative feedback on each aspect was counted. Using a binary scale for learner feedback (positive: +1, negative: -1), mean numbers were calculated per aspect (row-wise means, used below to quantify the quality of different aspects). As responses were given in freeform to open questions, not all aspects were covered in each interview. Moreover, the binary scale does not indicate how strong the learners' opinions were. The relative numbers of positive and negative opinions nevertheless indicate strong and weak aspects of the interaction.

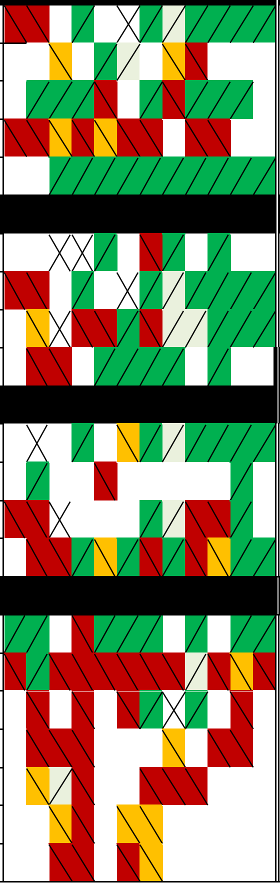
Language learning: The most positive responses were that participants would like to practice again with Furhat (+1.7) and that a majority stated that the conversation was good for learning (+0.8). Feedback was discussed by nine participants who correctly noticed that learner errors were not corrected. Four participants found this to be positive, as they became more relaxed than speaking with a human teacher who would correct them, while five wanted feedback and felt that the robot was not paying enough attention (-0.1). The level of the robot's utterances was adequate (+0.8) but many criticised that the robot did not clarify its utterances when the participants did not understand (-1.3).

Social context: The robot's personality (+0.3), the role as an L1 speaking peer (+0.8) and its ability to reply to within-topic questions (+0.5) were positively received, but it was criticised for not responding to more general questions.

Conversational interaction: The variation and duration of topics was adequate (+1.0), but the robot sometimes changed topic too abruptly (-0.3). Involving both learners was successful with some pairs, but not with all (+0.2). Regarding response times, negative learners either found that the robot took too long to respond, or, on the contrary, that the robot sometimes responded unnaturally quickly, without having to reflect before answering.

Robot moderation: The robot was perceived as understanding participants (+1.2), but the TTS speech rate was criticised for being too high (-1.3). Another problem concerned mispronunciations of participant names and unusual words, which could be confusing or detrimental for learning correct pronunciation. Opinions were divided regarding voice naturalness (-0.3): five learners found the synthetic speech good, while nine were negative, since it made understanding more difficult. Six learners mentioned that Furhat's responses to participant input were repetitive (in particular "*Interesting!*"). The robot's visual behaviour suffered from lack of emotions (-1.0), eye contact (-0.7) and natural-looking skin (-0.6).

Table 2. Post-session interview questions (reordered thematically), common responses and heatmap summarising responses per learner pair (sorted from negative to positive). Diagonal up signifies that one (light green) or both (green) learners were positive, diagonal down that one (orange) or both (red) learners were negative, crossed diagonals that they disagreed and empty cells that no feedback was provided. The last two columns show mean and standard deviation for each aspect.



Interview questions	Common responses	Learner pairs	μ	σ
Language learning				
Was the robot conversation good for learning?			0.8	0.8
	Giving feedback		-0.1	0.7
Was the conversation adapted to your level of Swedish?	Difficulty		0.8	0.9
	Clarifying		-1.3	0.2
Would you use Furhat to practice?			1.7	0.0
Social context				
How was the robot personality?			0.3	0.8
Are there benefits of robot moderators?	Less intimidating		0.8	0.8
What was your impression of Furhat?			0.3	0.9
Did the robot answer you?			0.5	1.0
Conversational interaction				
How did you like the conversation?	Variation of topics		1	0.6
How was the conversation flow?	Involving learners		0.2	1.2
Did the robot switch subjects abruptly?			-0.3	0.9
Were there unnatural pauses?			0	1.0
Robot moderation				
Did the robot understand you?			1.2	0.7
Was it easy to understand Furhat?	Speech rate		-1.3	0.7
What areas are the most important to improve upon?	Voice naturalness		-0.3	1.0
	Emotion display		-1	0.2
	Eye contact, blinking		-0.7	0.6
How lifelike was the robot?	Head-only set-up		-0.4	0.3
	Skin naturalness		-0.6	0.3

6.2. Assessment of the robot-led conversation by L2 Swedish teachers

In the next evaluation, 27 teachers of L2 Swedish and language café moderators (same respondents as for the pre-study questionnaire) rated the robot's interaction. Rating was done in two steps with a 5-point Likert scale, using the four areas of our analytical framework and statements in the upper part of Figure 2. First, respondents graded the importance of 1–4 aspects within each category (1 = unimportant, 5 = essential). In the second stage, they viewed a video of one robot-led conversation and rated how well the statements described the robot's performance (1 = strongly agree, 5 = strongly disagree). The results are summarised per area below, with ^{ns}/*/** respectively indicating if the difference between the two ratings was non-significant or significant at $p<0.05$ or $p<0.01$, using a non-parametric Mann-Whitney U-test.

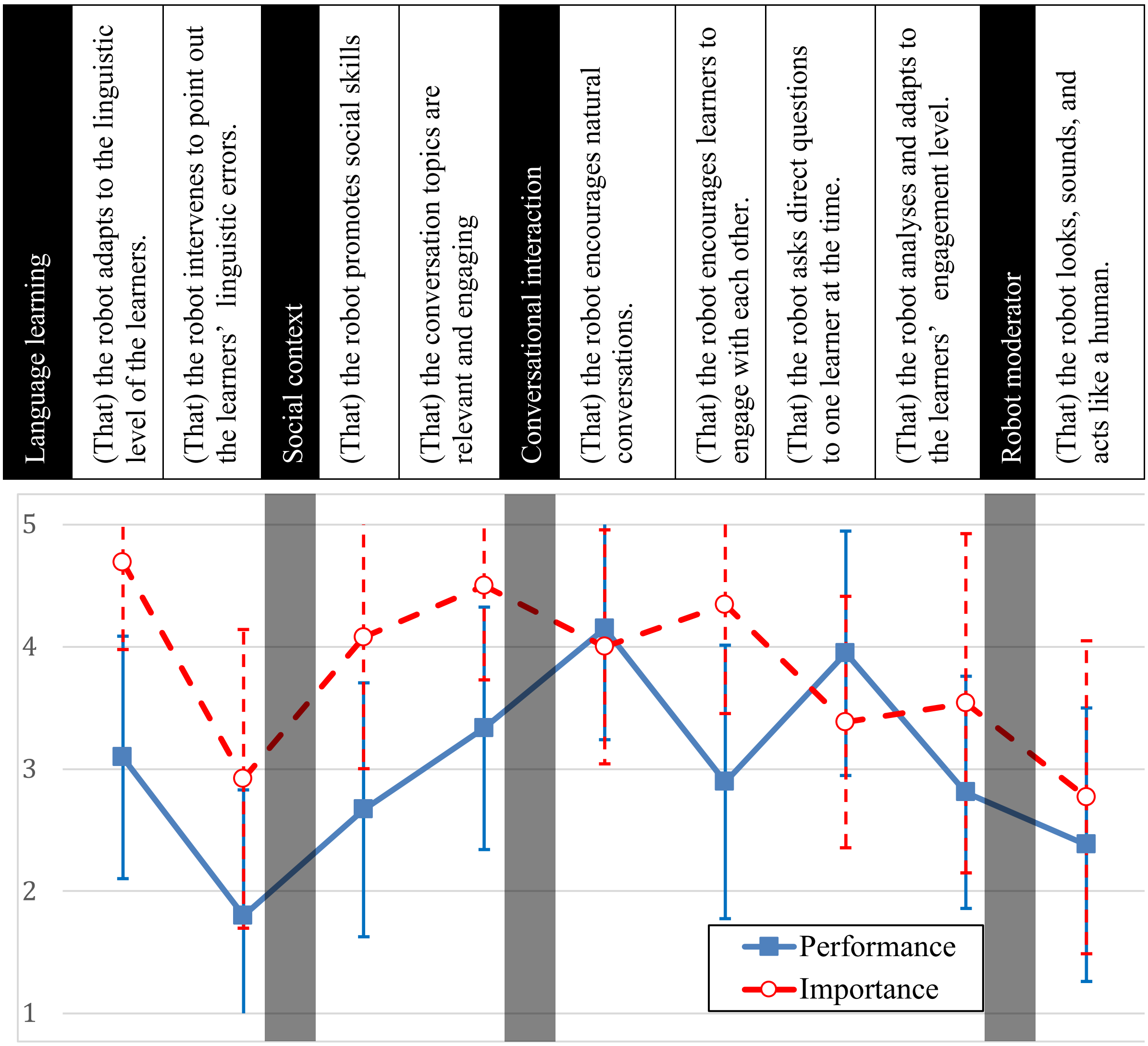


Figure 2. L2 teachers' description of the current robot behaviour (solid blue line; excluding 'That' in the statement) of different aspects and their importance (dashed red line), based on the statements above the graph. Error bars show standard deviation.

Language learning: The robot did not adapt to the learners' level as much as the teachers requested (performance: 3.1 *vs.* importance: 4.7**), but the fact that the robot did not correct linguistic errors was perceived as less problematic, since correction was viewed as less crucial (1.8 vs. 2.9**).

Social context: Topics were assessed as moderately relevant and engaging, which should be improved, as this is important (3.3 *vs.* 4.5**). The same applies to promoting social skills, where the gap is also large (2.7 *vs.* 4.1**).

Conversational interaction: The interaction was found to be natural (4.2 *vs.* 4.0^{ns}), but the robot should encourage the learners to interact with each other more (2.9 *vs.* 4.3**) rather than interviewing one learner at the time (3.9 *vs.* 3.4^{ns}) and should adapt more to the learners' engagement (2.8 *vs.* 3.5*).

Robot moderation: The robot was not assessed to sound and look like a human, but this was also not an important feature (2.4 *vs.* 2.7^{ns}).

In the free-form answers in the questionnaire, 12 teachers stated that the interaction was appropriate, regarding e.g., the type of questions and the robot's speech quality, 3 were more hesitant and 3 were negative. The criticism focused on how human-like and personal the robot was, its pronunciation and speech rate, the level of difficulty of some questions and the robot's inability to handle comprehension problems or to adapt to unforeseen conversation.

The teachers were further remarkably divided regarding if the current implementation would be useful in actual conversation practice, with 8 stating that it was already almost acceptable, 7 that it could be useful in some situations and 7 that it was far from ready.

6.3. Interaction analysis of robot-LC conversations

In the final level of analysis in this study, the 12 video-recorded sessions were examined to identify challenges, assess how they are currently handled and suggest improvements. The observations were made through inductive coding of interaction problems in the videos, illustrated in the extract in Appendix 1.

Language learning: The level of the practice overall seems appropriate for most, albeit not all, learners. The robot sometimes had to repeat questions, but this appears to be linked to speech rate and pronunciation of specific words, rather than to too complex utterances, since the learners were eventually mostly able to respond after the robot had repeated the question. If repetition did not help, a different robot utterance that changes the topic was used, but a more suitable strategy would be to lower the TTS speed or rephrase the utterance.

Lack of learner responses may also be caused by difficulties formulating an answer. Linguistic help to find the correct words is common in human-led conversation practice but is challenging for a robot and the most realistic would be to encourage the peer to assist.

Social context: Using participant names, countries, native languages, may improve the interaction, but as shown in Appendix 1, it introduces problems if the robot's information about learner placement is incorrect. Updating the robot's information or removing personalisation of robot utterances if a mismatch is detected would remediate the problem. The use of predetermined robot utterances for each state was found to limit the robot's ability to reply to personal questions.

Conversational interaction: Turn distribution is important to promote equal interaction from the learners, since one learner may be much more active than the other due to language level or personality. At present, there is no automatic tracking of the share of the conversation for each of the two learners, but it has been shown (Gillet *et al.*, 2021) that the robot may use gaze to achieve more even distribution.

The robot's topic transitions are currently explicit (e.g., "*Let's talk about something else.*"), whereas human conversations more seamlessly transition from one topic to another. A data-driven approach trained on human conversations could result in more natural topic changes.

Robot moderation: Lack of appropriate robot responses, or difficulties interpreting learner utterances, may be handled by involving the other learner (e.g., "*What do you think about this?*"). A particular case is learner input in another language, which would require the peer to assist, if the ASR is unable to recognise the intent from a partial transcription.

As the acknowledging feedback from the robot was found to be repetitive, the set of verbal and visual acknowledgement needs to be expanded from the five fixed responses currently used.

7. Discussion and conclusions

We conclude the article by discussing the study compared to the state-of-the-art, future work, its limitations, and its contributions to the CALL field.

7.1 Progress against state-of-the-art and future work

The present study extends previous work by demonstrating the possibility of using robots for more complex learning material (free social conversations, rather than practicing basic vocabulary), a more advanced learner group (intermediate level adults, rather than children) and a more demanding interaction with two learners (rather than one-to-one interaction or robot-to-whole-group). According to the evaluation of the robot-led conversations, several common expectations (such as varying topics, having an appropriate level, involving both learners and providing valuable, less intimidating practice) are at least partially met. However, others need to be addressed further for smooth conversation, as discussed below.

Language learning: Adaptation to learner level was highlighted in Section 6.2 as the aspect for which the difference between its importance and the robot performance was the largest and we will attempt adapting to learner level during sessions by using different sets of conversation topics and utterance complexity for different proficiency levels. A lower-level set will be used if a need is detected through verbal (e.g., many requests for clarification and repetition) or visual cues of confusion (Cumbal *et al.,* 2020) and a higher one if the learners are more proficient (e.g., produce more words per minute or longer utterances).

Feedback requires more attention, both from a pedagogical and a technological perspective. Pedagogically, opinions of respondents were divided regarding whether the robot should provide corrective feedback, and if so, how much. Technologically, it is currently beyond the state-of-the-art to robustly detect participant errors in free L2 conversations and provide adequate explicit feedback. We will instead explore whether the same robot responses could be used both for confirmation and as recast of incorrect formulations (e.g., *"So you **have been** in Sweden for a long time"* as response to both "*I have been in Sweden for three years*" and "*I be in Sweden three years"*).

Social context: The focus on societal content, which was highlighted in Section 3 as being important for adult learners, is quite unique compared to previous work in robot-assisted language learning. The robot provides facts about Sweden and asks the participants quiz questions about the country, but a more complex discussion about culture and society would have been beyond the capacity of most participants in these conversations. Such interaction would be interesting to explore with more proficient learners, but this requires that the robot is able to automatically create new utterances using methods for information-extraction-based question generation (e.g, Duan *et al.*, 2017) instead of manually created utterances.

The robot's personality was mostly rated positively by the learners, which is promising, since adult learners generally have higher expectations regarding social abilities than children, who accept stereotypic, toy-like robots (Engwall & Lopes, 2020). However, the robot's personality can be enhanced further by enabling the use of all robot utterances in the database at every stage of the conversation instead of only when the learner question had been anticipated. Another improvement would be that the robot remembers and integrates participant answers into ongoing conversations, using a learner database (e.g., Veerporten *et al.*, 2009).

Conversational interaction: The importance of the robot's role to monitor and distribute conversation turns when one learner dominates is clearly illustrated in Appendix 1. Ongoing work on determining the engagement level of the speaker, using textual and acoustic analysis of vocal features, and the non-active listener, using video analysis of facial expressions, indicates that such information is valuable to decide when a topic or addressee shift is needed.

Tests with state-of-the-art ASR (Cumbal et al., 2021) on robot-led conversation recordings show important challenges for ASR to correctly transcribe spontaneous non-native speech, but our current work also demonstrates that the impact of these problems may be minimised by selecting the most probable next robot utterance, given statistics of what utterance the wizard chose in previous, similar conversations. We will explore employing follow-up utterances that do not rely on fully understanding the learner input and, in case of low ASR reliability, avoid problems by involving the peer learner to continue the conversation.

Robot moderation: Future work is required on TTS quality, which received many negative comments both in terms of naturalness and intelligibility. To improve naturalness, which is important in L2 practice for understanding and as learning targets (e.g., In & Han, 2015), we will investigate switching to a deep learning TTS, since development over the past few years has led to remarkable improvement in naturalness. To improve intelligibility, the TTS speaking rate will be lowered for more difficult utterances. To determine which these are, a data-driven approach will be used, so that an utterance that is not understood by a learner will not only be repeated at slower speed for this learner but will also have reduced speed in subsequent conversations with other learners.

Eye contact and emotion display were not implemented in the present study and were two of the aspects that received the most negative comments. Matching gaze has been demonstrated in an earlier study using Furhat with an eye-tracker (Zhang *et al.*, 2017), and since the robot's face is computer animated, visual display of emotions and feedback can readily be added if robot utterances are manually or automatically coded for sentiment. Both eye contact and emotion display will be studied further in our future work.

7.2 Limitations

A major limitation is that the study intentionally does not evaluate the learning effectiveness of the practice, and this needs to be investigated. The studied interaction is moreover short (10 minutes), made once for a limited number (24) and one category (adults at intermediate level) of learners. The results are hence diagnostic with the aim of providing guidelines on how to improve this, and similar, L2 conversation practice. Future evaluations, after addressing aspects identified in this study, should study longer, repeated learner-robot interactions with a fully autonomous robot. Using repeated sessions is important, since learner preferences for the robot's interaction may change over time (Engwall *et al.*, 2020). Interaction with an autonomous robot needs to be investigated, since it may be different compared to with a wizard-of-Oz setup. This study presents conversation practice in Swedish, but in almost all respects, pedagogical requirements and analysis of interaction aspects would be similar regardless of language (except that English TTS and ASR may be more robust).

7.3 Contributions of the study

The study explored four important areas for development of conversational practice that to large extents are valid also for other types of conversation practice, led by a computer-animated intelligent tutor in CALL software or even a human moderator, except for aspects purely related to [robot] technology. Learners and teachers highlight the importance of using topics that are familiar and relevant, of establishing a personal relationship between the moderator and the learners, and of the moderator being active and adaptive to different learners, regarding moderating the conversation, its level of complexity and feedback given to learners. These requirements are natural for a human moderator but require additional efforts to fulfil with educational robots and intelligent tutors in CALL. This study has shown that some requirements for robot-led conversation practice are rather close to being fulfilled, such as motivating adult learners to engage in spoken practice and defining a suitable robot interaction strategy based on a social interviewing scheme inspired by language cafés. Other aspects, such as automatically adapting the linguistic complexity and the topics to the level of individual learners or providing corrective feedback and pedagogical support, will be more difficult to achieve. However, with additional development, robot-led conversation practice could serve as a complement to human-led lessons already with the current state-of-the-art, since the learners found the practice to be valuable, engaging and less intimidating than speaking with a human moderator.

Acknowledgements

This work was supported by the Swedish Research Council under Grant 2016-03698.

Author contributions

The article was written by the first three authors, partly based on material (interviews and surveys in sections 3 and 6.2, data collection in section 6.1) gathered by authors 4–11 as part of their BSc theses supervised by the first author.

Ethical statement

The authors have no conflicts of interest. All participants in the experiment signed an informed consent form.

References

Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B. (2012). Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction. In Esposito A., Esposito A.M., Vinciarelli A., Hoffmann R. & Müller V.C. (eds) *Cognitive Behavioural Systems. Lecture Notes in Computer Science*, 7403, 114-130. https://doi.org/10.1007/978-3-642-34584-5_9

Baxter, P., Ashurst, E., Read, R., Kennedy, J. & Belpaeme, T. (2017) Robot education peers in a situated primary school study: personalisation promotes child learning. *PLoS One* 12(5):e0178126. https://doi.org/10.1371/journal.pone.0178126

Belpaeme, T., Vogt, P. *et al.* (2018). Guidelines for Designing Social Robots as Second Language Tutors. *International Journal of Social Robotics*, 10: 325–341. https://doi.org/10.1007/s12369-018-0467-6

Cumbal, R., Lopes, J. & Engwall, O. (2020). "Detection of Listener Uncertainty in Robot-Led Second Language Conversation Practice," in *ICMI '20: International Conference on Multimodal Interaction*, 625–629. https://doi.org/10.1007/s12369-018-0467-6

Cumbal, R., Moell, B., Lopes, J. & Engwall, O. (2021) "You don't understand me!": Comparing ASR Results for L1 and L2 Speakers of Swedish. *Interspeech*, 4463-4467. https://doi.org/10.21437/Interspeech.2021-2140

Duan, N., Tang, D., Chen, P. & Zhou, M. (2017). Question Generation for Question Answering. Proceedings of Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/D17-1090

Engwall, O., Lopes, J. & Åhlund, A. (2020). Robot Interaction Styles for Conversation Practice in Second Language Learning. *International Journal of Social Robotics*, 13: 251–276. https://doi.org/10.1007/s12369-020-00635-y

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

Gillet, S., Cumbal, R., Pereira, A., Lopes, J., Engwall, O. & Leite, I. (2021). Robot Gaze Can Mediate Participation Imbalance in Groups with Different Skill Levels. *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 303–311. https://doi.org/10.1145/3434073.3444670

Hautopp, H. & Hanghøj, T. (2014), Game Based Language Learning for Bilingual Adults. *8th European Conference on Game-Based Learning*, 191-198.

Hong, Z.-W., Huang, Y.-M., Hsu, M., Shen, W.-W. (2016). Authoring robot-assisted instructional materials for improving learning performance and motivation in EFL classrooms. Educational Technology & Society, 19: 337–349.

In J. & Han, J. (2015). The Acoustic-Phonetics Change of English Learners in Robot Assisted Learning. *Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, 39-40. https://doi.org/10.1145/2701973.2702003

Johnson, L.W. & Valente, A (2008). Tactical Language and Culture Training Systems: Using Artificial Intelligence to Teach Foreign Languages and Cultures. *AAAI*, 1632-1639.

Khalifa, A., Kato, T., Yamamoto, S. (2017). Measuring Effect of Repetitive Queries and Implicit Learning with Joining-in Type Robot Assisted Language Learning System. In

ISCA workshop on Speech and Language Technology in Education, 13–17. https://doi.org/10.21437/SLaTE.2017-3

Kirschner, P., van Merriënboer, J. (2013) Do Learners Really Know Best? Urban Legends in Education, Educational Psychologist, 48(3): 169-183. https://doi.org/10.1080/00461520.2013.804395

Lee, S., Nohn H., Lee, J., Lee, K., Lee, GG., Sagong, S. & Kim, M. (2011) On the effectiveness of robot-assisted language learning. *ReCALL* 23(01):25–58. https://doi.org/10.1017/S0958344010000273

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

Neri, A., Cucchiarini, C. & Strik, H. (2002). Feedback in Computer Assisted Pronunciation Training: Technology push or demand pull? *International Conference on Spoken Language Processing*, 1209-1212.

Randall, N. (2019). A Survey of Robot-Assisted Language Learning (RALL). *ACM Transactions on Human-Robot Interaction*. 9(1): Article 7. https://doi.org/10.1145/3345506

Shin, J. & Shin, D-H. (2015). Robot as a Facilitator in Language Conversation Class. *Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, 11-12. https://doi.org/10.1145/2701973.2702062

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

Verpoorten, D., Glahn, C., Kravcik, M., Ternier, S., & Specht, M. (2009). Personalisation of Learning in Virtual Learning Environments. In U. Cress, V. Dimitrova & M. Specht (Eds.), *Lecture Notes in Computer Sciences. Learning in the Synergy of Multiple Disciplines,* 52-66. Berlin, Germany, Springer-Verlag. https://doi.org/10.1007/978-3-642-04636-0_7

Wang, Y. H., Young, S. & Jang, J. (2013). Using Tangible Companions for Enhancing Learning English Conversation. *Educational Technology & Society*, 16 (2), 296–309.

Westlund, KJM., Martinez, M., Archie, M., Das, M. & Breazeal, C. (2016) Effects of framing a robot as a social agent or as a machine on children’s social behaviour. *IEEE international symposium on robot and human interactive communication*, 688–693.

Wik, P., & Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. *Speech Communication*, 51(10): 1024-1037. https://doi.org/10.1016/j.specom.2009.05.006

Zhang, Y. Beskow, J. & Kjellström. H. (2017). Look but Don’t Stare: Mutual Gaze Interaction in Social Robots, *9th International Conference on Social Robotics,* 556-566. https://doi.org/10.1007/978-3-319-70022-9_55

About the authors

Olov Engwall is Professor in Speech Communication at KTH. He received his PhD from KTH in 2002 and has since led several research projects on computer-animated tutors, for which he received the Christian Benoît Award in 2004, and robot-assisted language learning. He has chaired several international workshops on the use of speech technology in language learning.

José David Lopes is Research Associate at Heriot-Watt University. He has a PhD from Instituto Superior Técnico, Lisbon, Portugal and has worked with spoken dialogue systems as a post-doctoral researcher at KTH.

Ronald Cumbal is PhD student at KTH under the supervision of Engwall and Lopes.

Gustav Berndtson, Ruben Lindström, Emilie Jin, Ella Johnston, Patrik Ekman, Eric Hartmanis, Gara Tahir, Michael Mekonnen are MSc students of the Industrial Engineering and Management programme, with a specialization in Computer Science.

Author #1 ORCiD. Olov Engwall, https://orcid.org/0000-0003-4532-014X
Author #2 ORCiD. José David Lopes, https://orcid.org/0000-0002-8773-9216
Author #3 ORCiD. Ronald Cumbal, https://orcid.org/0000-0003-4472-4732

Appendix 1

L1	Robot	L2
	Hello and welcome. My name is Furhat, what is your name?	
My name is L1		My name is L2.
	Nice to meet you, L1 and L2!	
Likewise.	[HT0] Where do you come from, L1?	
I am from Morocco.		
	How long have you been learning Swedish, L1?	
I have been three years in Västerås, but I was ill and. I moved to Morocco for long time. About three years.	Mmm.	
And came back. 10 months been.		
	You speak Swedish very well already!	
What?	Which languages do you speak?	
I speak Arabic, Swedish, English.		
	[HT1] And you, L2, which do you know?	That’s me.
	[HT2] And you, L2, which do you know?	I can [pause] I speak Azerbaijan [pause]
	Do you feel that any of those languages help when learning Swedish?	[silent]
	Can you tell us about something you find difficult when learning Swedish?	… Sorry?
	Can you tell us about something you find difficult when learning Swedish?	
		Ah. Difficult is for me… Pronunciation. And grammar. New… New word.
	What do you find difficult? [HT1]	
I think. It is difficult in Swedish language. I think.		
For me: listening. Listening.	Mhm.	
It is difficult for me when I heard, when I heard persons… talk… fast. Sometimes I cannot understand well.	Interesting.	
Yes. When I listen. When I to write and to read, I, no problem.		
	Interesting. [HT2] But what do you think…	[silent]
But I have no problem with grammar. [HT1] I have studied English, so I think many things in English they exist in Swedish language. And many words there is in Swedish language.	Mhm.	
	Interesting.	
Yes.	I was going to ask about something totally different. [HT2]	
	What country do you come from, L1? [HT1]	
What country? From Morocco. I come from Morocco. I tell you [Laughs]		
	Tell me more.	

Many times come from Morocco. From Fes, [*in French* Cultural capital]. The head in Morocco. Cultural capital in Morocco. There are many universities. [*In English* old] old universities. Institutes. The head. Cultural head in Morocco.

Interesting.

Yes. [HT1, HT2] What do you miss most from Morocco?

What? What do you miss most from Morocco?

I do not hear well. What do you miss most from Morocco?

I do not know one word. If I moved from Sweden, I would miss the snow.

[HT2] And what do you miss most? [silent]

[HT1] Let us talk about something else. What do you think about Sweden? [HT0]

Ah. Sweden. First it is democratic country. And nice country. Good people. Life is not difficult. But for me, weather is first problem here. In Morocco there is 300 days about sun. We call Morocco Sun country. And when I come here, six months winter. [Laughs] Cold and snow. It is snow in Morocco, but [*in English* in part of] Morocco. In Morocco, it is many weather… only one season. Seasons, seasons, season.

Many think that it is cold. But not me since I have my fur hat.

Yes. What do you two think is the best in Sweden?

What? What do you two think is the best in Sweden?

For example. Life here is not difficult. It is easy.

Communications. And bus. Buses and the train and all. All. Mhm

[to L2] What is it called? I do not know

[undistinguishable] … [pause]

[HT1, HT2] How would you describe the

[overlapping] In Stockholm Swedish weather?

There are no problems of life here. Only the weather and what you want to do at work after. Get job.

[HT2] What do you think?

I continue to study Swedish. [L2 protests bodily that L1 steals his turn]

After I check with my coach. Works at the employment centre. I will check later. Not now. In a year maybe. But I need to talk and understand very well Swedish to get a job. Thank you!

[HT2] What do you think? I think that. I think that the best is in Sweden…

I don't like warm weather. Weather is nice for me. I

Mhm like Swedish weather.

Because. I come from Azerbaijan. It is very hot. I don't

Why? like… hot weather. That is why…

Mhm. … I am in Sweden.

Interesting. Thank you!

We need to stop now. Thank you for the chat.

Thank you! Thank you!