International Journal of Social Robotics (2022) 14:1067–1085
https://doi.org/10.1007/s12369-021-00849-8

Is a Wizard-of-Oz Required for Robot-Led Conversation Practice in a Second Language?

Olov Engwall[1] · José Lopes[2] · Ronald Cumbal[1]

Accepted: 19 November 2021 / Published online: 5 January 2022

Abstract
The large majority of previous work on human-robot conversations in a second language has been performed with a human wizard-of-Oz. The reasons are that automatic speech recognition of non-native conversational speech is considered to be unreliable and that the dialogue management task of selecting robot utterances that are adequate at a given turn is complex in social conversations. This study therefore investigates if robot-led conversation practice in a second language with pairs of adult learners could potentially be managed by an autonomous robot. We first investigate how correct and understandable transcriptions of second language learner utterances are when made by a state-of-the-art speech recogniser. We find both a relatively high word error rate (41%) and that a substantial share (42%) of the utterances are judged to be incomprehensible or only partially understandable by a human reader. We then evaluate how adequate the robot utterance selection is, when performed manually based on the speech recognition transcriptions or autonomously using (a) predefined sequences of robot utterances, (b) a general state-of-the-art language model that selects utterances based on learner input or the preceding robot utterance, or (c) a custom-made statistical method that is trained on observations of the wizard's choices in previous conversations. It is shown that adequate or at least acceptable robot utterances are selected by the human wizard in most cases (96%), even though the ASR transcriptions have a high word error rate. Further, the custom-made statistical method performs as well as manual selection of robot utterances based on ASR transcriptions. It was also found that the interaction strategy that the robot employed, which differed regarding how much the robot maintained the initiative in the conversation and if the focus of the conversation was on the robot or the learners, had marginal effects on the word error rate and understandability of the transcriptions but larger effects on the adequateness of the utterance selection. Autonomous robot-led conversations may hence work better with some robot interaction strategies.

Keywords Robot-assisted language learning · Conversational practice · Non-native speech recognition · Dialogue management for spoken human-robot interaction

1 Introduction

In many scientific studies [1–7], the interaction between an educational robot and second language (L2) learners is controlled by a wizard-of-Oz, i.e., a hidden human controller who selects the appropriate robot response given the current state of the interaction and the learner input. Some studies [2,8–11] have employed a fully autonomous robot recognising a limited word corpus, but in general it is considered that automatic speech recognition (ASR) is not robust enough for L2 learners and that utterance selection by an autonomous dialogue manager does not allow for adequate learning activities. The wizard-of-Oz setup is therefore employed to allow for human processing of the spoken input and decisions on the appropriate continuation of the interaction. However, a common long term goal is to achieve autonomous robot behaviour based on data from wizard-of-Oz-controlled experiments, e.g., by letting the wizard's choice of robot responses guide utterance selection for an autonomous robot in similar situations. The main aim of this study is to investigate how well

✉ Olov Engwall
engwall@kth.se

José Lopes
jd.lopes@hw.ac.uk

Ronald Cumbal
ronaldcg@kth.se

[1] KTH Royal Institute of Technology, Stockholm, Sweden

[2] Heriot-Watt University, Edinburgh, UK

this method works for conversational practice in a second language.

One step towards making robots autonomous is to reduce the information provided to the human wizard to correspond more closely to the input that would be available to an autonomous robot [12], in particular displaying ASR transcriptions of user utterances rather than giving access to the actual audio. The problems with speech recognition errors may be more severe in social conversations with L2 learners, since the social dialogue makes the learner input more difficult to predict. We therefore first investigate both the objective accuracy of the ASR transcriptions of L2 learner output and how understandable they are for a human wizard, who should use them to select the next robot utterance. We then explore how well the dialogue could continue, with either a wizard basing the decisions on ASR transcriptions or autonomous selection of robot utterances.

This work continues in the tradition of work performed on native speakers interacting in dialogues with imperfect speech recognition [13,14] and expands it to the case of L2 dialogues. Previous work on, respectively, L2 learners' interaction with educational robots, the reliability of ASR for L2 speakers and the use of dialogue data to guide the actions of an autonomous dialogue system are therefore first presented (Sect. 2). The set-up for the current spoken language practice with L2 speakers of Swedish is then described (Sect. 3), in particular in relation to robot utterance selection and the four different interaction strategies that the robot employed during the social conversations. The strategy influences how much the robot controls the interaction, which may have an important impact on both the accuracy of the ASR results and the difficulty of selecting adequate robot responses for learner utterances.

The motivations of the study are: (1) observations that ASR transcripts of L2 learners in conversations have a substantially lower accuracy than for L1 speakers (due to less correspondence with the ASR's acoustic and language models) or for task-based dialogues (due to less constrained input) and (2) that social dialogues have larger freedom, which signifies that many different robot responses may be adequate. In other words, it can be expected that the ASR accuracy is low for L2 learners in social conversations, but also that the interaction may continue successfully even with low ASR accuracy. The study hence investigates if these factors balance each other, in order to answer the overarching question if a wizard-of-Oz is required to control the robot in social conversations with L2 learners or if a sufficiently adequate interaction could be achieved using an autonomous robot.

Previously collected audio recordings of wizard-of-Oz-controlled robot-led conversations in L2 Swedish [15] were submitted to Google's Cloud Speech ASR in batch to get transcriptions of learner utterances and evaluate their correctness and intelligibility (Sect. 4). Their actual usefulness was further tested by letting the same wizard select robot utterances for the original conversations based on ASR transcriptions of learner utterances, rather than on the audio (Sect. 5). Three different autonomous utterance selection methods were then implemented and assessed with respect to how well they performed compared to manual selection by the wizard. Since the Word Error Rate (WER) for learner utterances has been found to be high [16], we investigate if the autonomous methods could be based on knowledge about what response the wizard choose in similar conversation contexts rather than on ASR transcriptions. This on the one hand discards information provided in the learner utterances, but on the other hand avoids problems arising from low accuracy ASR.

The situation can be compared to a noisy cocktail party, at which one may have large difficulties hearing what the collocutors say, but it is nevertheless often possible to respond adequately by employing social conventions and knowledge about the topic that is discussed. This strategy may be successful for L2 conversation practice at beginner to intermediate level, which has an underlying dialogue structure that makes interactions similar, and with the three-party setting, which has been shown to reduce problems linked to, e.g., sudden topic shifts [17].

The main research questions for the study are:

[R1] Does the adequateness of manually selected robot utterances in robot-led multiparty social conversations depend on how understandable the ASR transcriptions of L2 learner utterances are?

[R2] Can autonomous selection of robot responses, without information from previous learner utterances, match manual selection that uses ASR transcriptions of previous learner utterances?

The initial hypotheses (numbered to reflect their respective relations to the two research questions) are that: (H1.1) Transcriptions from a state-of-the-art ASR will have a high WER, which would make them inadequate as input to an autonomous dialogue manager. They will also have an important variability, in the sense that some speakers and utterances will be recognized with fairly high accuracy, whereas others may even be incomprehensible. (H1.2) The adequateness of robot utterances selected by a human operator will depend on how understandable the ASR transcription is, but also on the wizard's experience of similar conversations. (H2.1) Autonomous selection of robot utterances will be inferior to a human operator's, but the difference might not justify the use of a human wizard-of-Oz instead of an autonomous system. Finally, the robot's interaction strategy may influence (H1.3) the WER and the understandability of the ASR transcriptions, (H1.4) how important this understandability is for the manual selection of robot utterances and (H2.2) the feasibility of making the robot utterance selection autonomously.

2 Related Work

This section presents previous work in three different areas relevant to the present study: firstly, social interaction between robots and L2 learners; secondly, studies on how well automatic speech recognition is able to interpret free-form utterances by L2 learners; and thirdly, methods for autonomous robot response selection or generation in social human-robot interaction.

2.1 Robot-led L2 Conversation Practice

Robot-assisted language learning studies have in general been performed with children, and this is in particular the case for social interaction aimed at practising spoken conversations. The main reason for this may be that the robots' appearance and capabilities are suitable for children, whereas adult learners have expectations, such as natural and meaningful interactions, that may not be met by most available robots [18].

Robot-led conversation studies performed with children range from basic question–answer exchanges [19,20] and explicit practice of conversation elements [21,22], to task-driven role-play scenarios of specific situations [10,11,23] and freer interaction [5,8,24]. These studies have shown that robots can be a motivational factor that increases the learners' interest and decreases anxiety regarding speaking in the L2. Further that non-verbal expressions, such as facial and body gestures to express emotions are important for the robot's interaction, not the least since it can to some extent can compensate for shortcomings in the text-to-speech synthesis and ASR. Moreover that it is beneficial to let the learners interact or familiarise themselves with the robot together with peers, rather than alone; that social interaction strategies and forming personal relationships between the robot and the learners are important to maintain the learners' interest over time; and that it is often more appropriate to give the robot a role as the learners' peer, rather than as a more knowledgeable tutor. These aspects, discussed in more detail in recent reviews [18,25,26] are important considerations for the present work, despite differences between child and adult learners [18].

The work by Khalifa [7,27] on using a conversational setting with two tele-operated Nao robots and one adult learner is a rare example of robot-assisted language learning for adults and the closest to the present line of work. The main similarities are the three-party setting and the structure of the conversation, with one robot leading the interaction by asking both the other robot and the learner similar questions. The main differences are the single-learner setting and that the interaction sequence was fixed, so that the second robot was always addressed before the learner, since the aim of the study was to investigate the extent to which the learner picked up new expressions from the robot's responses and started to use these. The current setting with two learners introduces additional complexity in turn-taking and in peer interaction between the learners, in particular when the robot's interaction strategy is to transfer the initiative to the learners.

Due to the scarcity of work with adult learners, the main source of information for the present robot-led conversation practice instead comes from questionnaires and interviews with moderators and participants in language cafés [15,28]. These gatherings provide realistic and effective conversation practice for adult L2 learners using group conversations between one or several native speakers and a number of L2 learners, who exchange information, ideas and opinions on e.g., everyday life, the news and personal experiences. At language cafés open to larger communities it is often the case that participants do not know each other and have very different background and linguistic level. It is therefore often the case that native speakers have a role as conversation leaders with responsibilities to initiate suitable topics and distribute turns to encourage all participants to speak. The conversations tend to focus on similar sets of general topics on a rather superficial level (e.g., comparing home countries and languages, hobbies and personal matters). The motivation for the present line of work is to study to what extent a social robot can act as conversation leader in such a setting. To charter interaction strategies of human conversation leaders, a questionnaire was sent out to language café conversation leaders (106 respondents) [15] and semi-structured interviews were performed with 27 language café participants, two L2 Swedish teachers and two researchers respectively specialising on language cafés and L2 Swedish education [28]. Based on this information, four different robot interaction strategies and a set of conversation topics were defined, as described further in Sect. 3.2.

2.2 ASR Performance for Conversational L2 Utterances

Most previous work on ASR for L2 speech have focused on assessment of the learners' utterances and there is a significant body of studies on mispronunciation detection, automatic pronunciation assessment and sentence semantics scoring. However, these are less relevant for the present study, since the task of processing non-native conversational utterances is vastly different. The learner input is not known beforehand and linguistic assessment is not the primary target, but rather communication of information, which means that the focus is on transcribing the learner utterances correctly enough to be able to adequately continue the dialogue. Studies on ASR for L2 conversation utterances are rare, but it is well-known that it is a challenge, as both phonetics and semantics may differ from native standard.

A study in which 44 Japanese university students recorded 13 elicited imitation sentences [29] resulted in a 34.3% word

error rate (WER) when transcribed by the Google Speech API, compared to 10.6% WER for one American English and one British English speaker. The transcription errors for the native speakers were mainly caused by proper nouns, and unusual word combination, word order and linguistic structures. For the L2 speakers, the substantially higher WER was partly due to mispronunciations, but since predictive syntax algorithms are used by the ASR, it recognised certain words that were labeled as mispronounced by a human annotator, while misrecognising others that were labeled as correctly pronounced.

Another effort to improve L2 English ASR employed dual supervised learning of a Recurrent Neural Network-based ASR trained with recordings of Japanese and Polish speakers' English presentations on Youtube [30]. The results were 11–19% *character* error rates, but it should be noted that corresponding word error rates are usually 3–4 times larger [31] and that an estimated WER would therefore be 40–60% [31, Fig. 1].

A third study adapted a deep neural network (DNN) in Kaldi for multi-language speech recognition [32] to English, Italian and German, using sentences read by 72 children aged 9–10. The WER was 31.7–53.2% depending on the L1–L2 combination and if the DNN was trained with monolingual or multi-lingual data, compared to 2.1–10.4% for native speakers.

Our own recent comparison [16] of ASR transcriptions of L1 and L2 speakers of Swedish when they were either reading sentences or interacting in language café conversations such as the ones in this study shows similar results. The WER in different combinations of L1/L2 speakers, spoken material and ASRs (Google Cloud, Microsoft Azure and Huggingface Wav2vec2) was investigated. For Microsoft Azure, which showed the best results on the data, it was found that WER was much higher for L2 speakers than for L1 speakers for read sentences (41% *vs.* 11%) and for conversation utterances than for read sentences (36% *vs.* 11% for L1 speakers, 51% *vs.* 41% for L2 speakers). It was further found that the difference in WER between L1 and L2 speakers was much smaller in the conversation setting than for read sentences (for Google Cloud ASR, the difference disappeared altogether with a WER of 0.41–0.42 for both groups). For the present study, we hence expect to find a high word error rate.

2.3 Autonomous Robot Utterance Selection

In general, conversational systems are either task-based or chat-oriented. The latter is considered as more complex with respect to dialogue management, since task-based dialogues may to large extents follow a branched decision tree to reach the goal, whereas chat-oriented dialogues are less predictable and it is more important that system responses are socially adequate. Both data-driven (corpus-based or deep learning generative models) and rule-based (state-based, with defined alternative transitions) approaches have been used for utterance selection. The purely data-driven approaches tend to produce responses that are too general or out of context and rule-based approaches have therefore dominated [33], at least until recently, when chatbots trained on billions of words have been able to generate adequate responses to user utterances [34,35].

However, the current setting is more difficult than the text-based interaction for which response generation has been successful. Firstly, there is an additional complexity of the two learners interacting with each other as well as with the robot. In our previous work, this has been handled by a wizard-of-Oz who keep track of and respond to input from two different learners, but approaches have recently been presented to manage multi-party conversations autonomously [33]. Secondly, the main challenge is to select an adequate response even if the output from the ASR is unreliable. The task is similar to work performed already two decades ago for native speakers [13], but since ASR for L2 speakers is still a challenge and since L2 conversations may differ from L1 interactions, it is worthwhile to revisit the topic. For native speakers, it was found [13] that with a word accuracy rate (WAR) of 70% in the ASR transcriptions, the wizards requested full repetition (signifying no understanding), clarification of missing and erroneous words, and verification (partial understanding) in 25% of their utterances, with 41% of these being requests for full repetition. Since we for this study expect a substantially lower WAR [16], the utterance selection methods need to handle such levels.

For utterance selection, we investigate data-driven approaches on top of a state-based conversation structure, i.e., determining the most probable transitions between different pre-defined dialogue states based on the selection that the wizard-of-Oz made in a previous study [15]. The approach of using data from wizard-of-Oz-controlled or crowd-sourced dialogues as training material for a dialogue manager has been proposed earlier [36–38], but then focused on task-based dialogues and specific domains (e.g., restaurant bookings, information requests, emergency control robots on oil rigs). It has been found [39] that it is more challenging to use a data-driven approach for social chats. Previous works have used e.g., Long-Short-Term-Memory or Convolutional Neural Networks [36] or a Hybrid Code Network [38], whereas we explore Next Sentence Prediction in BERT (Bidirectional Encoder Representations from Transformers) [40] and a custom-made simpler approach. BERT selects the next robot utterance based on either transcriptions of the learner input or the preceding robot utterance, while the custom-made method is based on statistics of the wizard's selection in similar situations in previous dialogues. Using BERT with robot-only input or the custom-made method circumvents the complication of unreliable ASR transcriptions.

3 Robot-led L2 Conversation Practice

The present work is built upon audio recordings and log files from an earlier experiment in which the social robot Furhat (c.f. Fig. 1) conducted social conversation practice with pairs of L2 learners of Swedish [15]. Furhat [41] is an anthropomorphic robot head that displays realistic facial expressions, using a computer-animated face projected on a 3D-printed mask. Lip movements are automatically synchronised with text-to-speech synthesis, with a third-party state-of-the-art voice speech synthesis from Cereproc. The robot has a motor-server neck that allows the head to physically turn towards different participants.

A wizard-of-Oz used short-cut keys to select robot responses among up to 10 dynamically changing, topic-specific, utterances, and seven static, general responses (*"Yes", "No", "I don't really know", "Maybe", "Mm', "Mhm"* and repeating the previous robot utterance).

The implementation of the state-based conversation dialogue flow is described in [15]. For the understanding of the present work, three aspects need to be summarised: (1) the general concept of the state-based dialogue flow, (2) the robot's four interaction strategies and (3) the collected user data.

3.1 State-based Dialogue Flow for Conversations

The flow of the robot's utterances is governed by a dialogue tree, with different topics and possible transitions between these. Figure 1 illustrates a simplified (since the number of interconnections between the states has been greatly reduced) but realistic start of a conversation. From the introduction state (`intro`), the conversation starts with questions to one of the learners regarding her occupation (`what_do_you_do`), then either turns to the other learner with similar questions (`what_do_you_do_2`) or asks the same learner about plans and dreams about a future career (`future_work`). As shown by the loops with logic test, the latter two states may be visited in any order and one of them may be excluded, before transitioning to the topic of the learners' spare-time interests (`hobbies`). After an optional follow-up state (`hobbies_2`), the dialogue branches into three different topics (sports, movies and the use of robots in society), depending on the learner answers. The three branches are then followed until reuniting again in the topic `what_did_you_do_this_weekend`. The example further illustrates an interaction strategy choice in that the conversation may include narration about robots (in this case about football championships for robots, films with robots in them, how robots are used in education) or Swedish trivia (the 1912 Stockholm Olympic Games), or exclude these states for a more learner-focused interaction.

Table 1 Description of robot interaction strategies, positioned in terms of focus and initiative

	Learner focus	Robot focus
Learner initiative	*Facilitator*, aiming for learner-learner dialogue, with robot encouraging learners to suggest topics, elaborate their answers and comment on peer utterances.	
Balanced initiative	*Interlocutor*, balancing questions to one learner or both and encouraging the peers to comment on each other's input, but also providing own stories and opinions. The robot is more personal, by referring to learners, their home countries and native languages by name.	
Robot initiative	*Interviewer*, asking one learner a set of social questions at the time before turning to the next with similar questions.	*Narrator*, the robot performs a semi-monologue or self-centered dialogue about Sweden or robots.

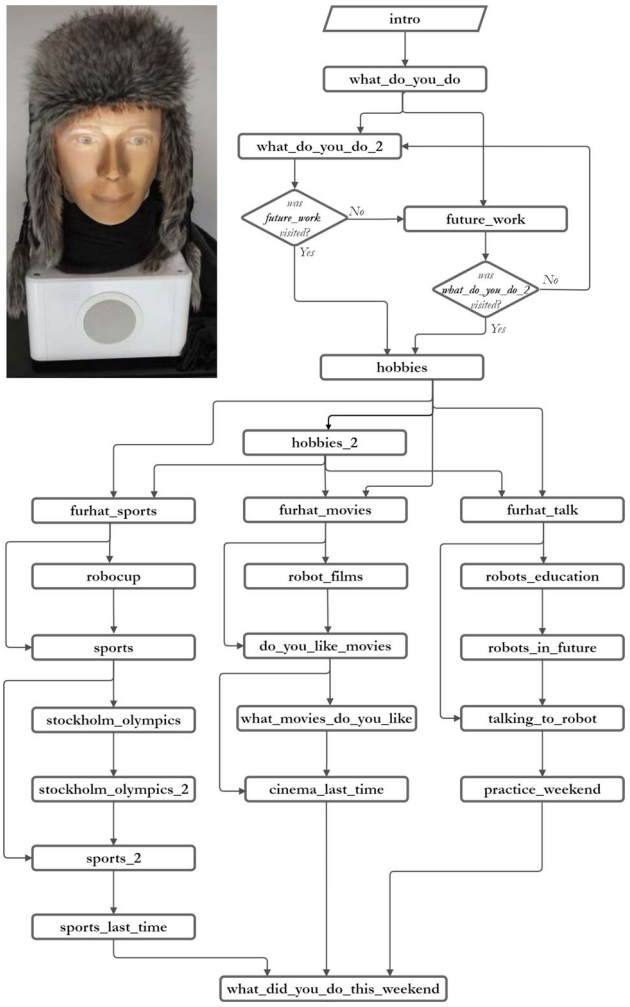


Fig. 1 The Furhat robot and a simplified states tree for the start of a conversation

The dialogue trees for the four robot strategies described in Sect. 3.2 have a similar structure, but since different states are visited with different strategies (e.g., the topics on robots and Swedish trivia may only be visited with Narrator and Interlocutor), the resulting dialogues become different. The main difference between the robot interaction strategies is, moreover, within the states, since the utterances are adapted to the strategy.

3.2 Robot Interaction Strategies

The four robot interaction strategies differ along the dimensions robot-learner initiative and robot-learner focus, as shown in Table 1, and in terms of robot utterances, as shown in Table 2. One strategy was maintained throughout the entire conversation and for the present study, the differences in initiative and focus are of interest since they may influence the ASR transcription accuracy. For example, learner answers to focused robot questions (for Interviewer) may be easier to

Table 2 Different possible utterances related to the same topic for the different robot interaction strategies

Facilitator	Interlocutor	Interviewer	Narrator
If you [two] compare your native languages; how similar are they?	What do you miss the most from Spain?	How long have you been in Sweden?	I have not been there, but I have traveled a lot. To conferences and exhibitions. I was part of an exhibition in London about robots in the future
Tell me more	Do you live close to that, Virginia?	Why did you come to Sweden?	Did you know that more than a third of the Swedish population moved to America between 1850 and 1930?
Can you [two] suggest something else to talk about?	If I moved from Sweden, I would miss the snow!	I am changing topic now	Have you heard about the ice hotel in Jukkasjärvi?

recognise than freer, long learner descriptions (for Facilitator) or very short learner feedback (for Narrator). Moreover, the robot strategies have different influence on the learners' affective state, and hence their vocal features such as loudness, speaking rate and intonation, which may in turn have consequences for the ASR. The differences between robot strategies may also influence how difficult selection of the following robot utterance is (e.g., it may be easier to correctly select a fitting utterance if the robot has the initiative in the dialogue (Interviewer) than when learners speak more freely (Facilitator).

3.3 User Study Data

The learners were 33 students (18 female, 15 male) in Swedish for Immigrant courses with varying background (from Afghanistan, Albania, Chile, Congo, Croatia, Cuba, Egypt, Eritrea, Iran, Iraq, Italy, Kazakhstan, Philippines, Poland, Somalia, Spain, Syria, Ukraine), age (20–52 years old, mean 32 years) and level of Swedish (they followed courses at B1 to B2 level, according to the Common European Framework of Reference, but differed substantially in actual level). In a pre-session briefing, which also included an informed consent form, the participants were informed that the goal was firstly for them to practice Swedish and secondly to assist in development of the robot, but they were not otherwise informed about the different robot strategies or instructed how to interact with the robot.

The study set-up was that each participant should have two conversations with one learner and two randomised robot strategies on the first day and two conversations with another learner and the remaining two interaction strategies on the second day. However, eight of these participants were not in class on day two and five new participants were therefore recruited to fill the gaps in the conversation pairs. This resulted in an imbalance both in the number of conversations that each learner participated in and the total number of conversations with each robot interaction style. 39 of the 50 conversations (i.e., 78 recordings of the entire conversation, one per learner, recorded with head-mounted microphones) were therefore selected randomly to get an, almost, equal number of conversations per robot interaction strategy, with 10 conversations for Interviewer, Narrator and Interlocutor and 9 for Facilitator.

The robot's dialogue logs—consisting of a summary of the current topic, the possible next robot utterances, the utterance that the wizard-of-Oz did select and the time stamp for the event—were also used. Temporally aligning the ASR transcriptions with the log files permits to replay the conversations in text format, with either manual or autonomous selection of the robot utterances at each turn.

Table 3 Word Error Rate (WER) and Ratio of utterances that Failed to be Recognized (RFR) for NU learner utterances in ND dialogues with different robot interaction strategies

	WER (%)	RFR (%)	ND	NU
Overall	41.6	11.5	29	1323
Facilitator	38.6	6.3	6	254
Interlocutor	41.7	8.1	7	284
Interviewer	46.3	5.9	9	422
Narrator	38.3	24.2	7	363

4 ASR Transcriptions of L2 Learner Utterances

The audio recordings were submitted to Google Cloud ASR for transcription of learner utterances on the word level. In the first experiment, we manually segmented audio recordings in order to process each learner utterance separately by the ASR. The large number of conversations (50) with 2052 learner utterances and a total of 27539 words (as compared to 348 utterances in 7 dialogues in [13]) signify that time-aligning and comparing the manual and ASR transcriptions is a very time-consuming undertaking. We therefore used a sub-set of 29 conversations, randomly selected from each of the four different robot interaction strategies to calculate WER. The number of substitutions, insertion and deletions of words in the ASR transcription compared to a manual transcription of the same utterance were counted and divided by the total number of words in the utterance. In addition, since the ASR has been found to fail to produce any transcription at all for some short (less than 4 words) and medium (5–10 words) utterances [16], the Ratio of samples that Failed to be Recognized (RFR) was also calculated, since these utterances were excluded from the WER calculation.

The results, overall and per robot strategy, presented in Table 3, indicate that the WER is high and comparable to previous work (c.f. Sect. 2.2). It should also be noted that the ratio of utterances that failed to be recognised (RFR) is particularly high with the Narrator strategy (four times higher than with Interviewer), since learner utterances with this strategy to a large extent consisted of short responses. The lower WER for Narrator hence only shows half the picture.

However, ASR transcriptions with high WER may still provide useful information for utterance selection if the important words are correctly recognised. On the other hand, transcription of separate utterances is a simplification of a real multi-party setting, in which ASR transcriptions of autonomously segmented utterances may include overlapping input from several speakers, which may make them more difficult to use to select a response.

To estimate how useful the ASR transcriptions are for utterance selection, the wizard from the original experiments

Table 4 Examples of utterances (translated from Swedish) classified into different categories of understandability. Bold font highlights the part of the utterance that was understandable in the context

Fully understandable	Interpretable	Partially understandable	Incomprehensible
"I like everything is adapted for wheelchair and it is pretty good", "Weekend I home clean and do laundry" **(D)**	"Because I can speak a bag and I watch [*in English:* British movies]" **(A)**	"cookies **music group favourite artist** trailer wagon" **(B)**	"GAIS [football team] colours in pastry", "Call mother" **(C)**
	"perhaps I don't talk about politics it is boring and there is little traffic" **(E)**	**"I like tennis and badminton** I don't know if it is penny Swedish shoes as well"	"fill in corresponding slightly shorter some message in first leg"

(A) Swedish *engelska [English]* being phonetically similar to *en väska [a bag]* and context make the utterance understandable. (B) Partly understandable when the topic is music preferences. (C) Linguistically correct, but out of context. (D) Fully correct transcription of a grammatically flawed learner utterance. (E) Second part understandable as "and that is a little boring", since Swedish *tråkigt [boring]* is phonetically similar to *trafik [traffic]*

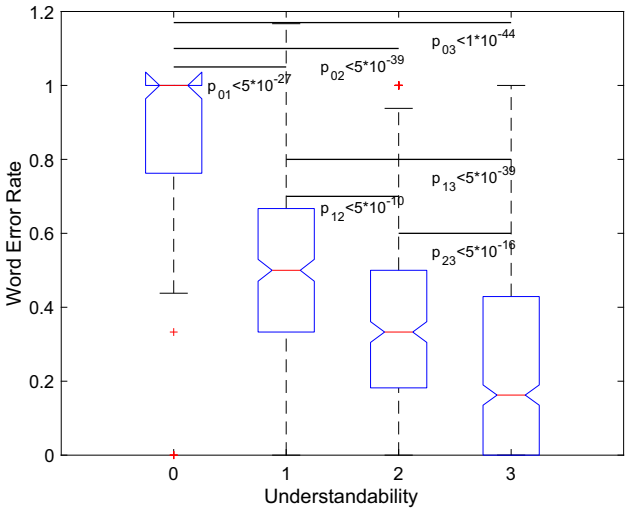
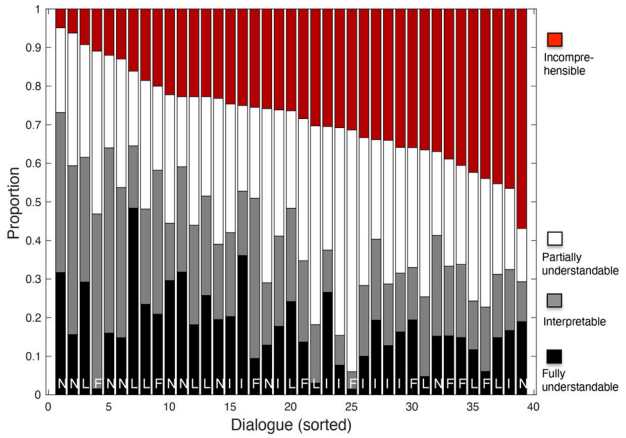


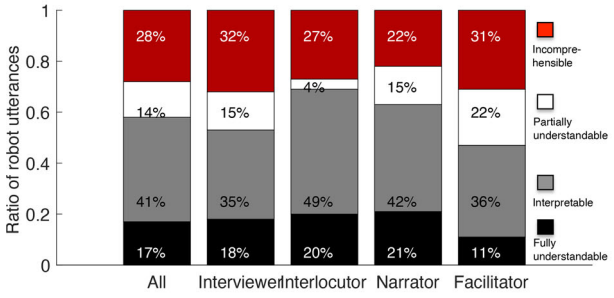
Fig. 2 Relation between the manual assessment of understandability and the WER. All differences between levels are significant

therefore assessed the understandability, both for the separate learner utterances used to calculate WER (to validate that the assessment of understandability was objective) and for utterances automatically segmented from speech recognition of the entire conversation (to investigate the realistic scenario). The rating of understandability used a four-level scale 0–3, with 3 being *fully understandable* and appropriate in the context, 2 indicating that the utterance was *interpretable* using context and/or linguistic knowledge, 1 signifying that the utterance was *partially understandable* and 0 being completely *incomprehensible* or out of context. Examples of utterances classified into the four categories are given in Table 4, which also illustrates why annotation by the original wizard was considered more appropriate than using several annotators making a more general assessment of the transcriptions' intelligibility. The assessment requires knowledge of phonetics (to interpret substituted words), of L2 learning of Swedish (to interpret common linguistic errors made by non-native speakers), of the topics discussed during the conversations (to interpret transcriptions that are partly understandable given the context or to know when they are out of context), and what information is required to select an adequate robot response.

For the assessment of objectivity, a Kruskal-Wallis test indicated that the WER of different understandability levels differ statistically ($p = 3.4 * 10^{-76}$). Using Mann-Whitney U-tests, highly significant differences in WER were found between all levels of understandability, as shown in Fig. 2. The mean WER for utterances judged to be fully understandable (M_3=0.25) was less than half than that for partially understandable ones (M_1=0.53) and less than a third of that for incomprehensible ones (M_0=0.86). It is hence confirmed that the wizard's annotations are closely related to



(a) Understandability per conversation (sorted in order of increasing proportion of Incomprehensible utterances). N: Narrator, L: Interlocutor, F: Facilitator, I: Interviewer indicate the robot interaction strategy.



(b) Understandability per robot interaction strategy

Fig. 3 Ratios of understandable ASR transcriptions

the objective WER. The interpretable utterances (M_2=0.43) have a much higher WER standard deviation (std_2=0.70 *vs.* std_0=0.25, std_1=0.43, std_3=0.32) since the annotator sometimes identified the actual meaning for utterances with high WER through phonetic knowledge of probable substitutions in the ASR transcriptions.

The understandability of the ASR transcriptions is summarised in Fig 3. It can first be observed that the proportion of Incomprehensible utterances per dialogue differs substantially, from less than 5% to 57%, as does the proportion of informative utterances (i.e., graded as 2 or 3), ranging from 6% to 73%. Hypothesis H1.1 is hence confirmed, since the WER of ASR transcriptions of L2 conversational speech is high and that many utterances are not understandable enough to guide selection of the next robot utterance. There is also an important variability between learner conversations.

Regarding the influence of robot interaction strategy it is found that Narrator has the lowest proportion of utterances assessed as Incomprehensible (0), while Interviewer has the highest, which is in line with the differences in WER reported in Table 3. Since a Kolmogorov-Smirnov goodness-of-fit test showed that the proportions of understandability for the different robot interaction styles were not normally distributed, a non-parametric Kruskal-Wallis significance test was used to investigate if there were significant difference between interaction strategies. However, no differences were found for fully understandable ($p = 0.17$), interpretable ($p = 0.16$) or incomprehensible ($p = 0.15$) utterances. Hypothesis 1.3, that the robot interaction strategy would influence the understandability of the ASR transcriptions of learner utterances, is hence not confirmed.

Both the WER and the manual assessment of understandability indicate that L2 conversational speech is still problematic for current state-of-the-art ASR. However, confidence scores for each transcription are provided by the Google Cloud ASR[1] and it may thus be investigated if this could be used to filter out unreliable transcriptions. Google provides no mathematical details for the calculation of the confidence score, but it is an estimate in the range 0.0–1.0 of the aggregated probability that each word in the transcription corresponds to actual words in the audio, and a higher confidence score hence signifies a higher probability that the utterance is correctly recognised.

As shown in Fig. 4, there is in fact a clear difference between the ASR confidence score means for utterances at different levels of understandability, and a single factor ANOVA with understandability as factor, indicate that there is a significant ($p < 1*10^{-42}$) difference between the means. This indicates that transcriptions with confidence score above 0.8 could provide important information to guide utterance selection.

Since a substantial proportion of the utterances were incomprehensible or only partly understandable, it is of interest to investigate if the nature of social conversations and the fact that the robot leads the conversations may signify that an adequate next utterance may be chosen based on conversation context and experience from previous similar conversations rather than learner input. Such an experiment for the current setting is presented in the next section.

5 Robot Utterance Selection

To evaluate robot utterance selection, we use an approach related to the method proposed in [42]: instead of conducting new user tests, recordings from the previous test are employed to replay the conversations and estimate how well utterances could have been selected autonomously, rather than by the wizard-of-Oz.

[1] https://cloud.google.com/speech-to-text/docs/basics#confidence-values.

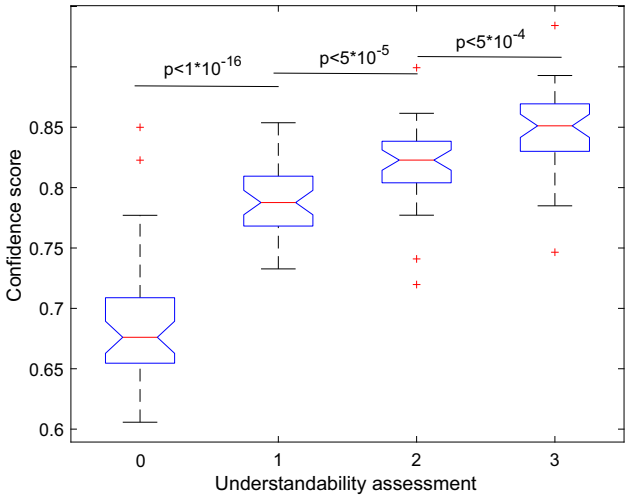


Fig. 4 Relation between manual assessment of understandability and ASR confidence score. All differences are significant

The robot's next utterance was selected using five different methods, which were compared with the choice made by the wizard-of-Oz in the original experiment:

Random selection (baseline) randomly picks one of the available robot utterances at each state.

Manual, ASR transcription-based, selection (benchmark) by a human operator, who selects the most appropriate robot utterance among the ones available for that state, just as in the original experiment, but basing the decision on ASR transcriptions, rather than the actual acoustics. The operator was the same wizard-of-Oz to provide a probable, albeit not theoretical, upper limit for the correspondence with the original selections (considered as gold standard).

Pre-defined selection is based on a manually created database of robot utterance pairs (U_1, U_2). When a specific robot utterance U_1 occurs in the conversation, a database look-up is performed to retrieve the following utterance U_2. This hence signifies that the robot always continues with the same utterance U_2 after U_1, regardless of the learner input. The creation of the utterance pairs was based on what would in general be suitable utterance U_2 after U_1.

Statistical selection based on language model, henceforth referred to as the "Language Model" method, employs the Next Sentence Prediction (NSP) of BERT [40] to determine the most probable next robot utterance given either a learner utterance or the preceding robot utterance. With the first alternative, the selection is made based on the manual transcription of what the learner actually said in the preceding turn. This should give an estimate of the best possible performance when using learner utterances as input in the current setting and with the training data available, since it corresponds to perfect transcription by the ASR. With the second alternative, it is instead considered that the transcriptions are too unreliable to use as input and utterance selection is instead performed by training on the sequences of robot utterances only.

We use the Swedish BERT model trained by the National Library of Sweden, KB-BERT [43]. The model is cased and pre-trained with the same hyperparameters as in the original BERT release. Fine-tuning the model for NSP was performed through the Transformer python library from Hugging Face. NSP consists in the binary classification of if the second sentence in a test sentence, pair is truly, the next sentence or a random one, and training and test set data were composed of 50% being the next utterance and 50% being a random one. Following the structure of the NSP task, all robot utterances were considered to be single sentences within a single conversation and, for the alternative when learner utterances were used, all learner utterances occurring between two consecutive robot utterances were joined into a single utterance.

The conversations were divided in training, validation and testing sets and the fined-tuned models were tested for NSP with a cross-validation procedure, where 1–3 conversations (depending on the dataset size) were randomly removed from the training data to be used for testing on never seen data. The NSP tests were repeated five times and the performance metrics were averaged over all experiments, achieving mediocre results. The prediction accuracy was 57.0% for learner utterance input in 26 dialogues with manually transcribed learner utterances (1941 utterances) and 61.1% for robot utterance input in all 39 dialogues (1695 utterances). When the robot strategies were separated for both training and test (in order to investigate if strategy-specific models could improve the results for the alternative with only robot utterances as input), the accuracy decreased to 56.2% (in 10 Interlocutor conversations), 55.3% (in 10 Interviewer conversations), 54.9% (in 10 Narrator conversations) and 53.1% (in 9 Facilitator conversations), most probably due to the decreased amount of fine-tuning training data.

The low prediction accuracy is explained by social conversations in general allowing many different combinations of utterances, and in particular by the robot responses being created so that they should be possible to combine in different order.

For the task of utterance selection, which is of interest in this work, BERT's probability for two sentences to be consecutive is used to rank the possible alternative utterances. The utterance with the highest probability that it follows the present one is selected.

Statistical selection trained on previous wizard choices, henceforth referred to as the "Statistical" method, is a custom-made method based on the utterance frequencies in the other conversations with the same robot interaction strategy (i.e., 10 conversations for Interviewer, 13 for Narrator, 12 for Interlocutor and 8 for Facilitator), but with other learners than in the present dialogue, hence ensuring that there is no overlap of conversations with the same learners between

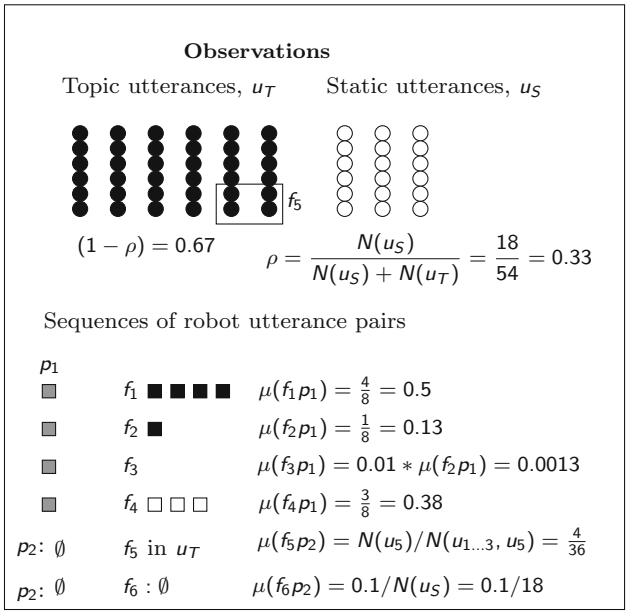


Fig. 5 Illustration of the custom-made statistical utterance selection method through a worked example. Top box illustrates the calculation of ρ from frequency observations of dynamically changing topic utterances u_T and static u_S. $\mu(fp)$ depends on if the utterance p occurred in the training set (p_1) or not (p_2) and if the following utterance f occurred after p in the training set ($f_{1,2,4}$) or not ($f_{3,6}$). N signifies number of observations. Bottom box exemplifies the calculation of the selection probability depending on if f is from the changing topic set of u_T ($f_{1,2,3,5}$) or the set of static utterances u_s ($f_{4,6}$). f_2 is a topic change utterance and f_4 is a repetition of the previous utterance

training and test data. The method disregards the learner input and instead uses implicit information of other robot-learner exchanges, by considering how often the wizard chose a certain utterance in a similar context. It is hence similar to the language model method with robot utterance input, but is of interest firstly since it is computationally cheaper than employing a large-scale language model and secondly since it might achieve better results by being specifically created for the task, in particular since the available data for fine-tuning a language model is very limited. The method is described below and illustrated in Fig. 5.

The statistical method consists of determining *observation probabilities* of utterance frequencies for *topic-specific and static* utterances in other conversations and *selection* based on these frequencies.

Observations probabilities $\mu(fp)$ of utterance f following the present utterance p are determined for the sets of dynamically changing topic utterances u_T ($f_{1,2,3,5}$ in Fig. 5) and seven static utterances u_S ($f_{4,6}$ in Fig. 5). For an utterance f that never occurred after utterance p in the other conversations with the same robot strategy (f_3 in Fig. 5), the observation probability is empirically set to 0.01 of the smallest observation probability $\mu(fp)$ of the other utterances f (i.e., a small, but non-zero probability).

If $\mu(fp)$ is undefined, because utterance p does not occur in the training set consisting of the other dialogues with the same robot strategy (p_2 in Fig. 5), the different $\mu(f_xp)$ for the utterances f_x that could occur after utterance p are instead set to the total relative frequency of utterance f_x in all other conversations (i.e., following any utterance). The observation probabilities for static and dynamic utterances are calculated separately for each set (f_5 *vs.* f_6 in Fig. 5). If an utterance f_x did not occur at all in the other dialogues the number of observations is set to 0.1 (f_6 in Fig. 5) to introduce a low, but non-zero, probability of selecting it.

Topic-specific and static utterances need separate probability weights, since the same static utterances are available at every robot turn, while the topic utterances change, which signifies that the relative probability of selecting one of the static utterances would be too high with equal weighting. The observed probability of choosing one of the static utterances (including repeat) ρ as opposed to one of the dynamically changing topic utterances (1-ρ) is therefore calculated, as shown in the upper part of Fig. 5.

Selection A weighted random sampling is used to select the following robot utterance f among the sets of static, with probability $\rho * \mu(fp)$, and dynamic topic utterances, with probability $(1-\rho) * \mu(fp)$.

The selection is adjusted for two cases. Firstly, it was observed in initial tests that explicit topic change utterances (e.g., *"Let's talk about something else"*, *"I was going to ask about another thing"*) were over-represented and in particular introduced too early topic shifts if utterance p came earlier in the present conversation than in other conversations. The reason for this over-representation is that topic changes occur as options both in several different topics and several times within each topic. To counteract too early topic changes, the number of turns n_{tc} that the topic change utterance has been an alternative within the topic is therefore used to gradually increase the probability of a topic change, as $\mu(tcp) = \frac{\mu(tcp)_{OBS}}{4-n_{tc}}$ for $n_{tc} \leq 3$. This means that the probability of selecting an explicit topic change the first time it is presented is one third of the observed frequency $\mu(tcp)_{OBS}$ and half the second time before reaching the observed probability.

Secondly, a factor n_r is introduced to handle repetitions of robot utterances, since the statistical method repeated them too often. The reason for this problem is that different learners in the original conversations often requested repetitions or clarifications of the same robot utterances. For these utterances, the wizard therefore repeated the utterance. If the learners continued to request a clarification, the wizard sometimes repeated the utterance again, but often switched to another utterance to attempt continuing the conversation instead. The statistical method captures the correlation between more complex robot utterances and following learner clarification requests, but repeated the utterance several times in a row. The probability of repeating an utterance is therefore adjusted by counting how many times n_r the utterance has already been repeated and dividing the probability for another repetition by n_r.

5.1 Adequateness of Robot Utterance Selection

The adequateness of the utterance selection was assessed using both an objective measure of correspondence with the wizard's choice in the original conversations and a follow-up manual assessment of how adequate the robot utterance was in the context.

The objective measure consisted in determining how often the automated methods selected the actual robot utterance from the original conversations among their top-1, top-2 or top-3 utterances. This assessment is motivated by considering the original choices by the wizard as the gold standard and the measure hence shows the extent to which the automated methods make similar choices as the wizard. By definition, only top-1 is defined for the pre-defined method, which always selects the same utterance U_2 after U_1, and the manual selection. For random selection, one, two or three alternatives are selected among the available utterances.

The results of the objective measure, shown in Table 5, indicate that the statistical method selects the same utterance as the wizard in the original conversations almost twice as often as the pre-defined method, more than three times as often as the Language Model and five times as often as random selection. The ratios for top-2 and top-3 show that the statistical method indeed makes very similar choices as the gold standard (85% of the time, the wizard's choice is among the method's three selected utterances).

The reason for BERT performing substantially worse than the custom-made statistical method for this task is the small amount of fine-tuning data in combination with the similarity of the alternative utterances to select among. Pre-trained BERT models perform well in many downstream tasks, but fine-tuning them for the specific task often leads to important improvement of the performance. However, from previous experience, we have observed that fine-tuning requires at least 1,000 samples for good performance, whereas the data from each robot interaction strategy separately only amounts to 400–500 samples, with a total of 1706 utterances. Without adequate fine-tuning, BERT tends to loose the ability to distinguish between semantically similar utterances [38], which is problematic for the present task, since all alternative utterances should, by design, be possible as follow-up to the current robot utterance. This causes many of the alternative utterances to have very high and very similar probabilities (differences between alternatives typically being as small as 0.00001), thus resulting in a close to random selection among a sub-set of the available utterances.

Table 5 Correspondence between the wizard's original choice and the automated methods' selection of one, two or three utterances. The Language Model (BERT) uses either learner utterances or the previous robot utterance to select the next robot utterance

	Top-1	Top-2	Top-3
Statistical	0.53	0.76	0.85
Language Model, learner	0.19	0.30	0.40
Language Model, robot	0.16	0.29	0.41
Manual	0.46	–	–
Pre-defined	0.30	–	–
Random	0.10	0.20	0.28

When analysing the utterance selection by BERT, the following problems are observed:

Timing The language model often selects the same utterance as the wizard, but one turn too early (7.6% of the selections with learner input and 12.4% of the selections with robot only input) or one turn too late (14.2% of the selections with learner input and 23.5% of the selections with robot only input). In some cases, this shift in timing would not affect the conversation, but in most it breaks the sequential structure of initial and follow-up questions that the wizard used to achieve a logical conversational flow.

Repetitions The language model repeated the same utterance in consecutive turns more than four times as often as the wizard (26.4% of the utterances based on robot-only input and 28.8% of the ones based on learner input compared to 6.6% for the wizard). The language model hence over-interprets the presence of consecutive repetitions, related to the learners asking the robot to repeat its utterance.

Explicit topic change For robot-only input, topic change utterances (e.g., *"Let us talk about something else."*) were severely over-used, with 26.1% of the utterances suggesting a topic change, compared to 1.9% of the wizard's utterances and 1.4% when the language model used learner input.

Use of non-verbal feedback Acknowledgement signals such as "Mhm" and "Mm" are indeed important and frequent in conversations (11.8% of the wizard's selection), but they were over-used by BERT (19.8% of the utterances based on

Table 6 Proportion of utterances assessed as adequate (rated 2 or 3), acceptable (1) or inadequate (0) for the four selection methods.

	Manual	Statistical	Pre-defined	Random
Adequate	0.83	0.87	0.72	0.44
Acceptable	0.13	0.09	0.17	0.30
Inadequate	0.04	0.04	0.11	0.26

robot-only input and 40.4% of the utterances based on learner input).

Robot response In social conversations it is common that all parties ask questions, and since the robot often did not have prepared answers for the questions, the wizard in the original conversations then had to reply with *"I don't really know"* (2.1% of utterances) or *"Maybe"* (0.7% of utterances). With learner input, the language model always selects the former (3.2% of the utterances), thus signalling that the robot is more often unable to answer than with wizard control, and without learner input, neither of the utterances is selected, thus signalling that the robot ignores the learner question. This problem is discussed further for the statistical method in Sect. 5.2

The two alternative utterance selection methods using a language model thus have problems that are partly similar and partly different. Since the custom-made statistical method performs much better than the language model on the objective measures for this dataset, only the former is analysed further below, since analysing the language model's performance more than in the previous paragraphs would not provide additional insights.

Even if the wizard's original choices are considered as the gold standard, they are often not the only robot utterance that would fit in the context of the conversation. A manual assessment was therefore also performed, using a four-graded scale 0–3, with 3 signifying *same* utterance as in the original conversation (i.e., top-1 above), 2 an *equally good* choice, 1 an *acceptable* choice and 0 *inadequate*. For the assessment, each robot utterance selected by each of the methods was presented chronologically together with the learners' response to the preceding robot utterance, and the annotator—the wizard from the original conversations—judged how well the selected utterance fit in the conversation.

From Table 6 it can be observed that both the manual and the statistical selection methods perform rather well at selecting an appropriate robot utterance, with the statistical method even being slightly better (mean for adequate, i.e. same or equally good, being 0.87 *vs.* 0.83). The pre-defined, fixed selection is in most cases inferior and random selection is much worse.

A one-way ANOVA with utterance selection method as factor reveals that the manual and statistical selection methods are significantly better (measured as proportion of inadequate utterances) than the pre-defined method ($p < 5 * 10^{-7}$), which in turn is significantly better than the random ($p < 5 * 10^{-12}$). There was no statistically significant difference between the manual and the statistical selections ($p = 0.94$). Over all robot strategies, it hence seems possible to employ the statistical method instead of using a wizard-of-Oz.

Table 7 Significance test of the difference in inadequate selection between utterance selection methods for I=Interviewer, L=Interlocutor, N=Narrator and F=Facilitator. ns: non-significant, *: significant at p< 0.05, **: significant at p< 0.01 or lower

	Manual				Statistical				Pre-defined			
	I	L	N	F	I	L	N	F	I	L	N	F
Stat	ns	ns	ns	ns								
Pre-def	*	**	ns	**	*	**	*	*				
Rand	**	**	**	**	**	**	**	**	**	**	**	**

The method in the first row always performs better.

When considering utterance selection per robot interaction strategy, Fig. 6, the graphs show similar overall distributions with respect to utterance selection method. However, a two-way ANOVA, with utterance selection method and robot interaction strategy as factors revealed that there were significant differences between robot interaction strategy (p <0.0001) and the interaction with utterance selection method (p <0.05) when considering the proportion of inadequate robot utterances. Subsequent repeated one-way ANOVA for the proportion of inadequate utterances with different robot interaction strategies, with utterance selection method as factor, yields the significance pattern displayed in Table 7, showing that the statistical method is significantly better than the pre-defined for all robot interaction strategies and manual selection is better for all but Narrator. The results are similar if the proportion of adequate utterances are investigated.

For Interviewer and Interlocutor, the most interesting aspect is the difference in distribution between same (black stack in Fig. 6) and equally adequate (grey stack) utterances. The selections by statistical method indicate that the sequential Q&A with Interviewer varied more than the three-party interactions with Interlocutor, and those by the pre-defined that conversations with Interviewer to a much larger extent had a fixed sequential order.

Narrator is handled almost as well by the pre-defined method (but with wizard and the statistical methods being better), which is explained by the fact that the robot can often adequately continue with the same response following the same dialogue path, regardless of what the learners are saying, since the robot is generally behaving in a rather egocentric manner in this strategy.

Table 8 Number of occurrences of different types of robot utterance errors with different selection methods and with different robot interaction strategies (I=Interviewer, L=Interlocutor, N=Narrator, F=Facilitator) over all dialogues

	Manual				Statistical				Pre-defined				Random				All
	I	L	N	F	I	L	N	F	I	L	N	F	I	L	N	F	
Ignore user question	1	0	12	6	0	0	0	1	0	1	19	6	0	0	17	4	67
Unclear relation	0	0	1	0	6	5	5	3	0	0	0	0	0	0	0	0	20
Lack of information	8	8	6	8	6	1	3	4	14	15	12	11	12	15	16	10	147
Repetition	3	3	3	0	0	4	4	4	0	0	0	0	0	0	0	0	21
Contradiction	1	0	3	1	6	1	1	2	10	19	6	12	15	12	9	11	109
Violation of common sense	1	0	0	0	0	0	0	1	5	7	8	6	8	5	13	5	59
Topic-change error	0	0	1	0	1	0	3	0	0	0	2	3	59	51	94	20	236
Social error	1	0	1	3	2	0	1	1	3	4	0	8	3	10	15	16	68
All	15	11	27	18	20	11	17	16	35	46	47	46	97	94	154	66	

Random selection is here made among the dynamic utterances only.

Facilitator is in fact handled better by the statistical method than by the wizard, illustrating that this interaction strategy, in which the robot more often should try to respond to learner input, is vulnerable to ASR transcription problems. It is further shown that the pre-defined method is not suitable when the learners have more initiative in the conversation.

5.2 Analysis of Inadequate Utterance Selection

The different types of inadequate robot utterances were analysed for different selection methods and robot interaction styles, as summarised in Table 8, in order to better understand what types of problems the different automated methods have and hence which improvements are needed. The type of error was manually annotated, based on a taxonomy for chat-oriented dialogue systems [44], by the same annotator in a separate analysis after the assessment of utterance adequateness had been made for all conversations. The annotation used eight (out of the original 13 labels in [44]) and slightly modified the interpretation for two labels (Lack of information and Social error) in order to suit the current conversation practice, as described below. The rationale for the annotation was to select the error type judged to be the most disturbing and/or most representative of the situation, as an inadequate robot utterance could often be labeled with several error types. For example, if a learner asked the robot a question, but the robot made a topic change instead of replying, this was labeled as *Ignore user question*, since this was considered as a graver communication error. Which error is the most severe may differ between contexts, but as a general rule, the errors were considered to be the most problematic in the following, descending order: Repetition, Ignore user question, Unclear relation, Lack of information, Contradiction, Violation of common sense, Topic-change error and Social error.

Repetition of an utterance when the learners did not request it was a problem for the statistical method, similarly as for BERT, as described above. It also occurred with manual selection, when the ASR transcription did not show any learner response, which was interpreted as a need for repetition.

Ignore user question signifies that the robot did not respond to an explicit question from the learners. The manual method should be able to handle questions, but Table 8 shows that learner questions were nevertheless sometimes missed, as they were not correctly recognized by the ASR, in particular with Narrator (since they were more frequent in this strategy). The problem is shared with the pre-defined method, whereas the statistical method mostly avoids this problem based on statistics of when learner questions are likely to occur.

Unclear relation indicates that the robot response was unrelated to the preceding learner utterance. This error, which occurred with the statistical method, is caused by the frequency analysis of robot utterance tuples. In the original conversations, the learners did sometimes ask the robot questions, but the wizard often had to resort to general, non-informative answers (e.g., *"Maybe", "I don't really know"*), since a more specific answer was often not available among the utterance alternatives. This made these answers relatively frequent, which, as a consequence, caused the statistical method to erroneously select such an answer even if the learners had not asked a question.

Lack of information here denotes that the robot did not repeat an utterance when a learner requested this. Repetitions are by definition not handled by the pre-defined method. The statistical method had, on the contrary, learned which utterances were difficult for the learners and performed better than the manual method, due to the unreliability of the ASR transcriptions.

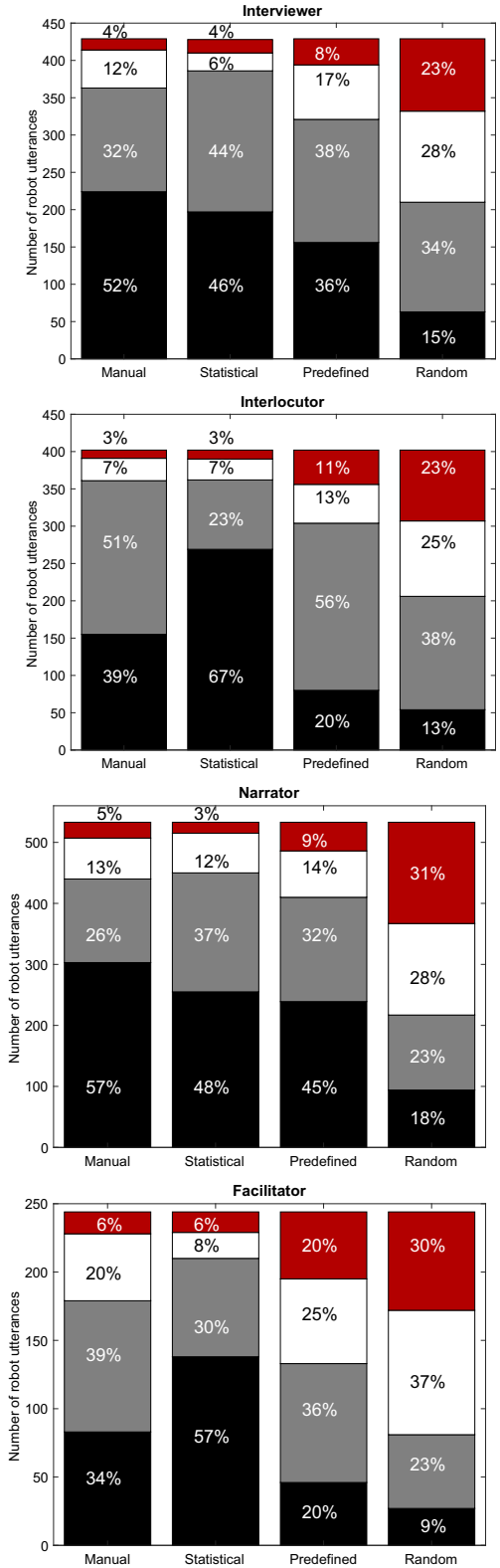


Fig. 6 Utterances assessed as, from bottom to top, identical to the original choice (black), equally good choice (grey), acceptable (white) and inadequate (red)

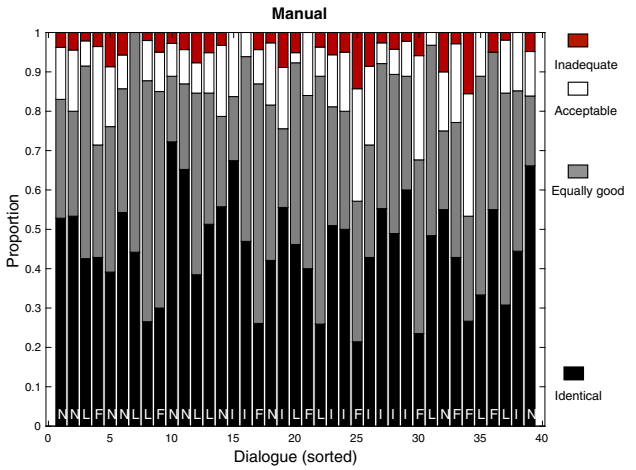


Fig. 7 Assessment of manually selected utterances compared to original wizard selection in each conversation, sorted in the same order as in Fig. 3a

Contradiction consists of different types of errors signalling that the robot had not understood what the learners were saying. It could be due to wrong assumptions about what the learner would answer; for example the sequence Robot: *"Do you like travelling?"*, Learner: *"No, I don't!"*, Robot: *"Why do you like travelling?"* or following up *"Why did you come to Sweden?"* with *"Oh, I am sorry to hear that."* when the learner provided a positive answer (e.g., to live with his wife, or to study ballet) or vice versa. It could also be that the robot asks a question that the learner has just provided the answer for by own initiative, or continuing as if the learners had replied to a robot question when they had not. Finally, it could be a contradiction between two consecutive robot utterances; for example *"I have been there!"* and *"I have not been there yet."* when the learners are talking about the same geographic place. The different types of follow-up errors occurred with different strategies. While the first types occurred predominantly in the Interviewer and Interlocutor strategies, when probing the learner for more information on a topic, the latter occurred with Narrator, and to some extent Interlocutor, when the robot should provide its own reactions to learner utterances. Except for Interviewer, the statistical method performed similarly to manual selection and substantially better than the pre-defined selection for this type of error.

Violation of common sense signifies utterances that were out of context in the conversation. This was not a problem for the manual and statistical selections, but were more frequent with pre-defined and random.

Topic-change error denotes that the robot explicitly suggested to change topic or abruptly changed topic too early or when the learner had said something that the robot should have responded to.

Social error was used for the special case that the robot did not appropriately acknowledge learner utterances using non-verbal feedback (*"Mm"* and *"Mhm"*). The statistical method was in fact at least as successful as the manual selection in determining when such non-verbal feedback would be an adequate response, whereas it was not included in the repertoire of the pre-defined, which consequently continued with a robot utterance that could be interpreted as the robot not listening to the learner.

When analysing the above types of errors and their frequencies it is worth noting that no adaptations could be made from the original conversation strategy set-up created for the wizard-of-Oz setting. As discussed further in Sect. 6, some of the errors could probably be reduced in frequency by changes in the robot interaction strategies, in the state-based utterance sequences or in the selection method.

5.3 Influence of ASR Accuracy on Manual Selection

It was further analysed whether the quality of the ASR transcriptions influenced the manual utterance selection. That is, if conversations with a higher ratio of fully understandable or interpretable ASR transcriptions also had a higher ratio of adequately selected robot utterances. Linear regression models were fit between, on the one hand, ratios of fully understandable, interpretable and incomprehensible ASR transcriptions or mean understandability and, on the other hand, ratios of identical, equally good and inadequate utterance selection or mean adequateness. The ratio of fully understandable ASR transcriptions *did* have a statistically significant positive impact on the combined ratio of identical and equally good selection ($\mu_{sel} = 0.76 + 0.38 * \rho_{ASR}$, $p = 0.017$) and the mean assessment of the utterance selection ($\mu_{sel} = 2.08 + 0.9 * \rho_{ASR}$, $p = 0.0078$). The combined ratio of fully understandable and interpretable transcriptions did not have any statistically significant impact on the ratio of identical or equally good selection ($p = 0.11$) or mean assessment ($p = 0.09$). Neither did the ratio of incomprehensible transcriptions lead to a higher ratio of inadequate utterance selection ($p = 0.77$) or a lower mean assessment ($p = 0.67$). This is in line with the fact that Fig. 7 does not show a decreasing adequateness left-to-right, which would otherwise be the case, since conversations are sorted with increasing ratio of incomprehensible transcriptions. Considering the robot interaction strategies separately, the relationship between the ratio of fully understandable transcriptions and mean adequateness of utterance selection was significantly positive for Narrator ($\mu_{sel} = 2.05 + 1.32 * \rho_{ASR}$, $p = 0.040$), but not for other strategies or measures.

In is hence found that both the ratio of fully understandable ASR transcriptions and how predictable the conversation sequences are determine how adequate manual selection of robot utterance can be.

6 Discussion: Limitations and Future Work

This section discusses the results of the study, in particular with respect to limitations and future work.

6.1 ASR Transcriptions

The above experiments show that the WER is high and the level of understandability rather low for ASR transcriptions of L2 learner utterances in robot-led multiparty conversations, with a large proportion of the transcriptions being either incomprehensible or only partially understandable. It was further found that the variation in understandability is large between different utterances, but that this to a large extent is captured by the ASR confidence score. This signifies that ASR transcriptions can, on the one hand, not be used without constraints, but that they, on the other hand, may provide adequate information when the confidence score is high. Note that since the ASR transcriptions were assessed after the utterance selection experiments (as discussed in Sect. 6.3 below), confidence scores were not used to guide utterance selection. Addressing a number of limitations in the current study with respect to ASR transcriptions could potentially increase their usefulness further.

Firstly, the ASR is based on the default language model and dictionary. There are possibilities to provide state-of-the-art ASR with a custom add-on lexicon to increase the probability to recognise words that are specific to the setting. Since social conversations are free by nature, the vocabulary may in principle be very diverse, but as there are a number of reoccurring topics during the conversations, such as comparing home countries and native languages or favourite music and movies, an add-on lexicon could avoid some misrecognitions.

Secondly, ASR is performed without providing context from the dialogue manager. Since the conversation is often led by the robot, learner utterances are frequently a direct response to the preceding robot utterance, which may be used to constrain the ASR to expect certain utterance formulations and/or specific vocabulary. Learners with different linguistic ability and personality may be more or less verbose in social conversations and these constraints may therefore not use a strictly defined language model, but providing a robot utterance-guided language model could nevertheless potentially improve the ASR results.

The above modifications could make ASR a viable input source for this setting, in particular in combination with improvements in utterance selection.

6.2 Utterance Selection Methods

The analysis of the adequateness of the selected robot utterances showed that the custom-made statistical selection

method performed as well as manual selection based on ASR transcriptions and clearly outperformed pre-defined utterance selection. Selecting robot utterances based on statistics of wizard choices in previous conversations, rather than on actual learner input, could hence be a viable strategy in this setting. Note that this is valid for this particular setting. In many other HRI situations, it is, on the contrary, essential to understand what the user is saying to respond correctly.

An even higher level of adequateness could probably be achieved by tailoring the conversations for autonomous robot utterance selection instead of wizard-of-Oz control; for example adapting the robot interaction strategies to guide (and consequently constrain) the learners more or taking advantage of the three-party setting to mitigate ASR transcription problems [17]. The system would then be less dependent on the learner input and it would be less vulnerable to unexpected learner utterances.

The statistical method could be improved by:

Firstly, including robot responses to frequently occurring learner questions, in general, and in particular within topics and/or robot interaction strategies where such questions are more frequent. That is, being able to answer the same type of questions that the robot asks the learners in each state, and being able to answer questions in interaction strategies that are more robot-focused (Narrator and Interlocutor). This would allow the robot to provide sensible answers to learner questions, which would in turn reduce the over-representation of general robot answers in the robot utterance pairs, thus preventing the Unclear relation answers identified in the present experiment.

Secondly, to reformulate follow-up robot utterances to not rely on the recognition of positive *vs.* negative learner responses (c.f. the example *"Why did you come to Sweden?"* above). While this has the downside of making the robot less personal and human-like (since the natural reaction would be to mirror the learners' positive or negative feelings in the robot response), it would in most cases be less detrimental for the conversation than an erroneous emotional display.

Thirdly, to reformulate follow-up questions to be more independent of the preceding robot question (and hence learner answer), in order to avoid communication breakdown when the robot wrongly assumes that/how the learner answered the previous question.

Fourthly, to make use of ASR transcriptions, when they are reliable. In the present study, the ASR transcriptions were discarded entirely, except for in the manual selection method.

6.3 Methodology and Assessment

Replaying previous interactions using modified processing and decision components can provide valuable insights into what works and what does not without having to involve new test subjects in the diagnostic evaluation step. This is an appealing alternative when the availability of users from the target group is limited or if practicalities (such as a pandemic) restrict the possibilities of conducting user tests. However, it will be essential to perform actual user tests to gather feedback from the users and to test interactions that do not need to follow the exact same path as in the original interactions. In the present study, replaying utterances in particular meant that all robot utterances and their connections needed to be the same as in the original experiment, which thus hindered improvements regarding robot utterances and interaction strategies.

All manual assessments in this study were made by one single annotator, and ratings of understandability and adequateness may hence have been subjective. However, the very clear relationship between the objective WER and the manual assessment of understandability shows that the latter may be considered to be largely objective. Moreover, as explained in Sect. 4, understandability is a specific measure to assess how informative the ASR transcriptions are for robot utterance selection, not of general ASR quality, and it is hence important that the annotator has specific knowledge about the utterance selection task. Similarly, the clear relationship between the manual assessment of how adequate the selection was and correspondence with the gold standard (the original selection being among the top1-3 choices) demonstrates that the manual assessment is in line with objective measures.

For the manual utterance selection, there is a potential risk that it was based on the wizard remembering the original conversations, rather than the ASR transcriptions. However, as the present experiment took place 3.5 years after the original conversations, this influence should be minor. Further, the influence of this bias would be to raise the benchmark and the statistical method nevertheless achieved similar results. It should be noted that even though the assessment of the understandability of the ASR transcriptions is described first (Sect. 4) as this is a more logical order of presentation, the experiment on replaying the robot-led conversations as a text-based interaction (Sect. 5) was in fact carried out before the ASR transcription assessment to avoid that the wizard's choice of robot utterances was influenced by previously assessing the ASR transcriptions.

7 Future Work & Conclusions

Hypothesis H1.1 is confirmed. ASR transcriptions of L2 social conversations have high WER (42%) and this make them difficult to understand for a human who should use them to select the next robot utterance. Further, there is an important variation in how understandable the transcriptions are, but it is found that the ASR confidence score could be used to identify transcription that are intelligible for a human reader.

Hypothesis H1.2 is partly confirmed. It was shown that fully understandable transcriptions led to more adequate manual selection of robot utterances, but also that the wizard could choose an adequate, or at least acceptable, robot utterances at almost all dialogue turns (96%) and that experience of similar conversations is hence also important.

Hypothesis H2.1 is rejected. The method based on statistics of the wizard's choices in similar conversations performs as well as manual selection, in terms of both objective correspondence with the original selection and subjective assessment of how adequate the choice was.

Hypotheses H1.3, H1.4 and H2.2 regarding the influence of robot interaction strategy are partly confirmed. WER was lower and understandability higher for the interaction strategy in which the robot is most active and the learners mostly provide direct and short reactions to the robot (Narrator) than for the strategy in which the robot requests the learners to provide the most personal information (Interviewer). However, these differences have marginal influence on how adequate the manually selected utterances are. Moreover, the autonomous utterance selection methods perform rather adequately for all robot interaction strategies. However, it should be noted that the statistical method was better than manual selection for the two robot interaction strategies in which the robot either drives the conversation with questions to one learner at the time (Interviewer) or, on the contrary, aims for encouraging learner-learner interaction (Facilitator).

While using a wizard-of-Oz for conversational L2 practice is still a safe and reasonable choice, the study demonstrates the possibility of achieving a similar interaction with autonomous selection of robot utterances. The natural next step is to perform a standard user test with an autonomous system to evaluate it with actual learners and get their feedback. This evaluation could potentially be set up as a modified Turing test, asking participants to identify if the robot was autonomous or remote-operated, in addition to providing general comments regarding improvements that could be made in the interaction.

Funding This study was funded by Swedish Research Council (grant 2016-03698 "Collaborative Robot Assisted Language Learning (CORALL)") and Marcus and Amalia Wallenberg foundation under grant MAW 2020.0052.

Declarations

Conflict of interest The authors declare that they have no conflict of interest. The original datasets analysed during the current study are not publicly available due to privacy issues, as the subjects have been guaranteed that their audio and video data will not be distributed. However, log files and ASR transcriptions are available from the corresponding author on reasonable request.

References

1. Zhen-Jia Y, Chi-Yuh S, Chih-Wei C, L, BJ, Gwo-Dong C (2006) A robot as a teaching assistant in an English class. In: Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06). pp 87–91
2. Park S, Han J, Kang B, Shin K (2011) Teaching assistant robot, robosem, in English class and practical issues for its diffusion. In: Proceedings of Advanced Robotics and its Social Impacts. pp 8–11
3. Tanaka F, Matsuzoe S (2012) Children teach a care-receiving robot to promote their learning: field experiments in a classroom for vocabulary learning. J Human Robot Interaction 1(1):78–95
4. Alemi M, Meghdari A, Basiri NM, Taheri A (2015) The effect of applying humanoid robots as teacher assistants to help iranian autistic pupils learn english as a foreign language. In: Social Robotics. ICSR 2015. Lecture Notes in Computer Science. Volume 9388
5. Mazzoni E, Benvenuti M (2015) A robot-partner for preschool children learning english using socio-cognitive conflict. Educ Technol Soc 18(4):474–485
6. Kennedy J, Baxter P, Senft E, Belpaeme T (2016) Social robot tutoring for child second language learning. In: The Eleventh ACM/IEEE International Conference on Human Robot Interation (HRI '16). pp 231–238
7. 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. pp 13–17
8. Kanda T, Sato R, Saiwaki N, Ishiguro H (2007) A two-month field trial in an elementary school for long-term human-robot interaction. IEEE Trans Rob 23(5):962–971
9. Han J, Jo M, Jones V, Jo J (2008) Comparative study on the educational use of home robots for children. JIPS 4(12):159–168
10. Mubin O, Shahid S, Bartneck C (2013) Robot assisted language learning through games: a comparison of two case studies. Austral J Intell Information Proc Syst. Vol. 13
11. Gordon G, Spaulding S, Westlund J, Lee J, Plummer L, Martinez M, Das M, Breazeal C (2016) Affective personalization of a social robot tutor for children's second language skills. In: Proceedings of AAAI Conference on Artificial Intelligence
12. Pereira A, Oertel C, Fermoselle, L, Mendelson J, Gustafson J (2020) Effects of different interaction contexts when evaluating gaze models in hri. 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI) pp 131–139
13. Zollo T (1999) A study of human dialogue strategies the presence of speech recognition errors. In: AAAI Technical Report FS-99-03
14. Cavazza M (2001) An empirical study of speech recognition errors in a task-oriented dialogue system. In: Proceedings of the Second SIGdial Workshop on Discourse and Dialogue. (09)

15. Engwall O, Lopes J, Åhlund A (2020) Robot interaction styles for conversation practice in second language learning. Int J Soc Robot 13:251–276
16. Cumbal R, Moell, B, Lopes, J, Engwall O (2021) You don't understand me! comparing asr results for L1 and L2 speakers of swedish. In: Interspeech
17. Arimoto T, Yoshikawa Y, Ishiguro H (2018) Multiple-robot conversational patterns for concealing incoherent responses. Int J Soc Robot 10:583–593
18. Engwall O, Lopes J (2020) Interaction and collaboration in robot-assisted language learning for adults. Computer Assisted Language Learning pp 1–37
19. You ZJ, Shen CY, Chang CW, Liu BJ, Chen GD (2006) A robot as a teaching assistant in an english class. In: Proceedings - Sixth International Conference on Advanced Learning Technologies, ICALT 2006. Volume 2006. (08) pp 87 – 91
20. Alemi M, Meghdari A, Ghazisaedy M (2015) The impact of social robotics on l2 learners' anxiety and attitude in english vocabulary acquisition. Int J Soc Robot 7(4):523–535
21. Wang Y, Young S, Jang JS (2013) Using tangible companions for enhancing learning english conversation. Educ Technol Soc 16(04):296–309
22. Wc Wu, Wang RJ, Chen NS (2013) Instructional design using an in-house built teaching assistant robot to enhance elementary school english-as-a-foreign-language learning. Interact Learn Environ 23(12):1–19
23. Lee S, Noh H, Lee J, Lee K, Lee G, Sagong S, Kim M (2013) On the effectiveness of robot-assisted language learning. ReCALL 23(1):25–58
24. Kanda T, Hirano T, Eaton D, Ishiguro H (2004) Interactive robots as social partners and peer tutors for children: a field trial. Human Comput Interact 19(1):61–84
25. van den Berghe R, Verhagen J, Oudgenoeg-Paz O, van der Ven S, Leseman P (2018) Social robots for language learning: a review. Rev Educ Res 89(2):259–295
26. Randall N (2019) A survey of robot-assisted language learning (rall). ACM Transact Human Robot Interact 9(12):1–36
27. Khalifa A, Kato T, Yamamoto S (2016) Joining-in-type humanoid robot assisted language learning system. In: Proceedings of LREC. pp 245–249
28. Engwall O, Lopes J, Cumbal R, Berndtson G, Lindström R, Jin E, Johnston E, Mekonnen M, Tahir G Learner and teacher perspectives on robot-led l2 conversation practice. ReCALL (Accepted)
29. Ashwell T, Elam JR (2017) How accurately can the google web speech api recognize and transcribe japanese l2 english learners' oral production? Jalt Call J 13(1):59–76
30. Radzikowski K, Nowak R, Wang L, Yoshie O (2019) Dual supervised learning for non-native speech recognition. EURASIP J Audio Speech Music Process 2019(1) 3:1–10
31. Lund W, Kennard D, Ringger E (2013) Combining multiple thresholding binarization values to improve ocr output. In: Document Recognition and Retrieval XX. (02)
32. Matassoni M, Gretter R, Falavigna D, Giuliani D (2018) Non-native children speech recognition through transfer learning. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp 6229–6233
33. Martinez VR, Kennedy J (2020) A multiparty chat-based dialogue system with concurrent conversation tracking and memory. In: Proceedings of the 2nd Conference on Conversational User Interfaces. CUI '20, Association for Computing Machinery
34. Adiwardana D, Luong MT, So DR, Hall J, Fiedel N, Thoppilan R, Yang Z, Kulshreshtha A, Nemade G, Lu Y, Le QV (2020) Towards a human-like open-domain chatbot. arXiv preprint arXiv:2001.09977
35. Roller S, Dinan E, Goyal N, Ju D, Williamson M, Liu Y, Xu J, Ott M, Shuster K, Smith EM, Boureau YL, Weston J (2021) Recipes for building an open-domain chatbot. In: EACL
36. Wen TH, Vandyke D, Mrkšić N, Gašić M, Rojas-Barahona LM, Su PH, Ultes S, Young S (2017) A network-based end-to-end trainable task-oriented dialogue system. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Valencia, Spain, Association for Computational Linguistics (April) 438–449
37. Budzianowski P, Wen TH, Tseng BH, Casanueva I, Ultes S, Ramadan O, Gašić M (2018) MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, Association for Computational Linguistics (October-November) pp 5016–5026
38. Lopes J, Garcia FJC, Hastie H (2020) The lab vs the crowd: An investigation into data quality for neural dialogue models. arXiv preprint arXiv:2012.03855
39. Jonell P, Fallgren P, Doğan FI, Lopes J, Wennberg U, Skantze G (2019) Crowdsourcing a self-evolving dialog graph. In: Proceedings of the 1st International Conference on Conversational User Interfaces. pp 1–8
40. Devlin J, Chang MW, Lee K, Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding
41. Al Moubayed S, Beskow J, Skantze G, Granstrom B (2012) Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In: Cognitive behavioural systems. Springer pp 114–130
42. Yamaguchi T, Inoue, K, Yoshino K, Takanashi K, Ward NG, Kawahara T (2015) Analysis and prediction of morphological patterns of backchannels for attentive listening agents
43. Malmsten M, Börjeson L, Haffenden C Playing with words at the national library of sweden – making a swedish bert. arxiv
44. Higashinaka R, Araki M, Tsukahara H, Mizukami M (2019) Improving taxonomy of errors in chat-oriented dialogue systems. In: D'Haro LF, Banchs RE, Li H (Eds) 9th International Workshop on Spoken Dialogue System Technology. Singapore, Springer Singapore, pp 331–343

Olov Engwall is professor in Speech Communication at KTH Royal Institute of Technology. He received his PhD in 2002 with a thesis on multimodal articulatory speech production modeling and has since focused on computer-animated tutors for pronunciation training and robot-assisted language learning. He currently leads a research project on Culturally Informed Robots in Learning Activities.

José Lopes is research associate at the Interaction lab at Heriot-Watt University. He received his PhD in 2013 from the Instituto Superior Técnico, Universidade de Lisboa, Portugal, and worked as a postdoctoral researcher at KTH Royal Institute of Technology 2014–2018. His main research topic is adaptive spoken dialogue systems.

Ronald Cumbal is a PhD student in robot-assisted language learning under the supervision of Engwall and Lopes.