International Journal of Social Robotics (2021) 13:251–276
https://doi.org/10.1007/s12369-020-00635-y

Robot Interaction Styles for Conversation Practice in Second Language Learning

Olov Engwall[1] · **José Lopes**[2] · **Anna Åhlund**[3]

Accepted: 18 February 2020 / Published online: 6 March 2020

Abstract
Four different interaction styles for the social robot Furhat acting as a host in spoken conversation practice with two simultaneous language learners have been developed, based on interaction styles of human moderators of language cafés. We first investigated, through a survey and recorded sessions of three-party language café style conversations, how the interaction styles of human moderators are influenced by different factors (e.g., the participants language level and familiarity). Using this knowledge, four distinct interaction styles were developed for the robot: sequentially asking one participant questions at the time (Interviewer); the robot speaking about itself, robots and Sweden or asking quiz questions about Sweden (Narrator); attempting to make the participants talk with each other (Facilitator); and trying to establish a three-party robot–learner–learner interaction with equal participation (Interlocutor). A user study with 32 participants, conversing in pairs with the robot, was carried out to investigate how the post-session ratings of the robot's behavior along different dimensions (e.g., the robot's conversational skills and friendliness, the value of practice) are influenced by the robot's interaction style and participant variables (e.g., level in the target language, gender, origin). The general findings were that Interviewer received the highest mean rating, but that different factors influenced the ratings substantially, indicating that the preference of individual participants needs to be anticipated in order to improve learner satisfaction with the practice. We conclude with a list of recommendations for robot-hosted conversation practice in a second language.

Keywords Robot-assisted language learning · Multi-party human–robot interaction · Collaborative language learning · conversational practice

1 Introduction

Conversation practice is fundamental when learning to speak a second language (L2). Practice can take place in the language classroom (but in this setting, individual learners get limited time to speak), in one-on-one sessions with a native-speaking tutor (which may not be a practical or economical alternative for most learners) or with native speakers in real-life situations (which may be too intimidating for beginner learners).

Language cafés provide an important complement by practicing real-life oral skills in a semi-structured and allowing setting. In its pure form, language cafés are an open gathering, where first language speakers and L2 learners meet to have social conversations in the target language. The focus is on conversation and communication, i.e., the content is given more importance than the form. Key concepts are that all interaction should be in the L2 language, that the topics should be universal enough to promote confidence and inclusion of all participants, and that all participants have a high acceptance for others' language errors, hesitations and slow interaction. Using this implicit contract, language cafés follow the concepts of communicative language teaching [1] and collaborative language learning [2] and employ a constructivistic view of learning.

In the original language café concept, the first language speaker's role is as an equal conversational partner, rather

✉ Olov Engwall
engwall@kth.se

José Lopes
jd.lopes@hw.ac.uk

Anna Åhlund
anna.ahlund@buv.su.se

1 KTH Royal Institute of Technology, Stockholm, Sweden

2 Heriot-Watt University, Edinburgh, UK

3 Stockholm University, Stockholm, Sweden

than as a teacher or conversation leader, even if first language speakers often take on additional responsibility to assist the L2 learners with linguistic problems. However, in communities with many immigrants, municipalities, libraries, NGOs or churches organize language cafés more like practice sessions for L2 learners, with one first language speaker as the person responsible. In this role, which we henceforth refer to as "moderator", the first language speaker prepares the session (choosing topics for the conversation, selecting a newspaper article to discuss, inviting a presenter to give an introduction, planning on how to divide the group and the available time etc) and also has a strategy for how to behave during the session in order to maximize the learning outcome for the participants.

A commonly encountered problem is that there is a shortage of first language speaker moderators, in general, and in municipalities with a large immigrant populations in particular. Technology-enhanced learning (TEL) solutions, and not the least social robots, can therefore provide an attractive alternative to give L2 learners more opportunities to practice spoken conversation.

Sweden is of particular interest in this aspect, since the number of immigrants has increased drastically in proportion to the total number of inhabitants over the last decade (the number of asylum seekers per year increased 10 times, from 16,303 in 2000 to 162,877 in 2015[1]) and some municipalities now have a majority of non-native speakers (an example is the town of Södertälje, where 53–57% in the age range 16–30 years—of the population of 92,700 had an immigrant background in 2017, after the arrival of 9132 new immigrants with resident's permit[1] since 2007[2]). This firstly leads to a high pressure on language education, with the number of adult participants in Swedish for Immigrants (SFI) courses having increased four times over the last 20 years (from 40,000 in 1997 to 73,000 in 2007 and over 163,000 in 2017[3]). Given that the number of SFI teachers has not increased at the same rate (from 2000 to 3600 full-time equivalents, of which only 35% are certificated[3]) there is a pressing need for complementary practice. Secondly, in communities with a large immigrant population, such as some areas of the above-mentioned Södertälje, which have an 80% immigrant population,[4] first language speakers of Swedish are in minority, which makes the language less spoken. Thirdly, as 86% of the Swedes report that they can hold a conversation in English,[5] it is extensively used as lingua franca in the Swedish society. However, mastery of Swedish is still important for integration in society in general and on the job market in particular.

To respond to this need for TEL in second language learning of Swedish, we are currently engaged in a research project on collaborative robot-assisted language learning (RALL), in which we investigate how a social anthropomorphic robot can be used in spoken conversation practice with pairs of L2 learners of Swedish. Language café moderators are suitable role-models for the robot, since the targeted spoken practice share the underlying pedagogical ideas of promoting fluency and self-confidence through non-judging conversations on familiar topics. Human–Robot interaction (HRI) could never fully replace human moderators, as it is extremely difficult to deal with the variety of the interaction that depends e.g., on the learners' L2 level, their background and if they already know the other participants. However, for conversation training on the basic to intermediate level, even human-led language cafés are often limited to a set of general topics focused on either comparisons between the home countries and languages, or personal matters, such as interests, family, food, or cultural preferences. Such conversations could in principle be moderated by a dialogue system following finite states in the interaction, as there are larger possibilities of generating robot utterances beforehand and of predicting user responses. An additional advantage of using robots in this setting is that previous studies have shown that they may reduce learner anxiety about making errors [3,4].

The present study has two parts. We firstly analyze, through a web-survey and observations, the strategies that human moderators employ in language café session, so that these strategies can guide the implementation of robot moderator strategies. We secondly explore how four implemented robot strategies are perceived by learners, using a semi-automated wizard-of-Oz controlled user test followed by a survey in which learners rate the conversation and the robot.

In order to approach the two goals set up above, we will first analyze the state-of-the-art in RALL focused on social interaction (Sect. 2), then investigate human moderator strategies, through a survey (Sect. 3.1) and observations of human moderator's behaviour in a set of lab-based three-party conversation practice sessions (Sect. 3.2), before defining strategies for a robot moderator (Sect. 4.1) and testing these in a user study (Sect. 5).

2 Collaborative Robot-Assisted Language Learning

Developing a set-up for a humanoid robot that can engage in a realistic social conversation with two L2 learners simultaneously is, to the authors' knowledge, unprecedented in Robot-Assisted Language Learning (RALL).

[1] Figures from the Swedish Migration Agency.

[2] Statistics from the municipality of Södertälje, "immigrant background" signifies that the person, or the person's both parents, are born in another country.

[3] Statistics from the Swedish National Agency for Education.

[4] According to Statistics Sweden.

[5] European Comission: Europeans and their Languages.

The set-up with one robot and two L2 learners is admittedly rather different from the traditional language café, with many learners, and possibly also more native speakers, but it is nevertheless judged to be sufficiently similar to make use of similar moderator strategies as the communicative approach and the conversation topics are similar. The reasons for choosing the current set-up, rather than other possible settings can be summarized as:

Robot with a larger learner group would be insurmountable for state-of-the-art speech technology, as multi-party dialogue management and more unconstrained L2 speech recognition would be required. In a three-party setting, the dialogues are more foreseeable and the problem therefore tractable.

Robot, native speaker and one learner would have the benefit of a native speaker providing linguistic support, but would make the robot role unclear and possibly superfluous, as we aim at investing the robot's potential as an independent conversation practice support for L2 speakers. In addition, the advantage of using robots to reduce learner anxiety about making mistakes in front of native speakers would be eliminated.

Robot and one learner would have the advantage of a focused practice for the individual learner, but would miss out on the important collaborative aspects, discussed next.

2.1 Benefits of Collaborative RALL

The HRI setting with two learners is taking advantage not only of the traditional benefits of collaborative language learning, in which learners learn from each other, but also in terms of increased robustness in the interaction with technology. Spoken interaction with a robot may fail when the automatic speech recognition (ASR) does not properly detect what a learner is saying, or when the learner does not understand the text-to-speech synthesis (TTS) generated robot utterance.

ASR for non-native speakers in a conversational setting is very challenging, since learner utterances may contain a substantial amount of errors of pronunciation, vocabulary and syntax. In such cases, the other learner in a collaborative setting can help to reformulate the input. However, it may also be that the ASR failed, despite the utterance being correct, because it was uttered with a foreign accent, on which the ASR had not been trained.

For the TTS, the problem may be that the learner does not know some words in a correctly produced robot utterance, in which case the peer may help with the understanding. However, it may instead be due to inadequate adaptation of the speech synthesis to L2 learners, resulting in robot utterances that are difficult to understand. In both cases when it is in fact the ASR or TTS that fails, it is important to have the support of the second learner to confirm that the problem lay with the robot, in order to avoid the impression that it was the learner who made a mistake. Such erroneous feedback from a TEL system may otherwise be detrimental for learning [5].

2.2 Related Work

RALL is in general a very new research field, with most work being carried out during the previous decade. A number of existing surveys [3,6–9] summarize earlier studies, as well as the general research questions and challenges. It has been known for quite some time that robots can be beneficial for language learning. In an early study [10], it was shown that Korean children practicing English during 40 min with the semi-humanoid robot IROBI were more interested in the learning activity than the groups practicing with web-based instructions or book and audio tape. As a consequence, the children in the robot group retained more of the material on the following day.

Previous work has almost exclusively targeted school and pre-school children, in particular from the perspective of increasing motivation for learning by employing robots in different roles. The robot can act as an assistant to a human teacher [4], as an independent tutor [11,12], as a peer learner [13] or a partner in solving a task [14], as an opponent in a game[15] or as a social companion [16,17].

In the scope of this work, and given that a comprehensive review of RALL [9] has recently been made, we concentrate on related studies where the robot acts as a peer or social companion, as they share the use of communicative language teaching and/or collaborative language learning with our study. We further focus on studies with more humanoid robots, rather than those using toy-like robots (e.g., Philips iCat and Lego Mindstorms in [8], the snowman Keepon in [18], the muppet-like Tega in [19]), since the focus is on developing practice that is suitable for adult learners and that is realistic, in its use of non-verbal social interaction signals.

The humanoid robot Robovie was employed as a social companion for learning English in Japanese classrooms over, respectively, 2 weeks and 2 months in two early studies [16,17]. The first [16] showed that the children's interest in interacting with the robot faltered after 1 week as they did not find the interaction stimulating over an extended period of time (Robovie could only remember 300 sentences and recognize 50 words, which limited the interaction). Consequently, for the second, longer, study [17] the robot's interaction behavior was improved by personalizing it to each user over time and expanding its social capabilities. Improvements in the children's level of English were found, and in addition, they responded that their main motivation for interacting with the robot was to form a friendship with it, illustrating the potentials of social interaction as a driving force in language learning.

Collaborative language learning was studied between the humanoid robot Mec Willy and Italian children aged 4–6 years [14]. The robot asked each child for help matching the English names for fruits and vegetables to the correct pictures and rather than giving correct answers it encouraged them to discuss their solution. This socio-cognitive conflict strategy is aimed at increasing the learners' awareness of potential differences in points of view and their ability to reason about their own. It was found that children retained substantially more words when learning together with the robot than with another child, showing the benefits of introducing a robot in a peer learning session.

Many recent RALL studies use humanoid Nao robots, e.g., Iranian children being taught English in whole-class setting, demonstrating that the robot decreased learner anxiety [4]; and European pre-school children in one-to-one interaction in practice sessions to learn English, Dutch or German, with a specific focus of increasing the rapport between the robot and the children [11].

Three studies with Nao robots are of particular interest for this study, as a peer or collaborative setting was used: 3–6 years old Japanese children were encouraged to teach English to a Nao robot in a verb learning game, by showing it how to perform the action [13]. The post-test indicated that children learned more words with the robot than without, hence demonstrating the dual benefit of collaborative language learning, as the peer who is acting as tutor also improves. Similarly, in another study, 9–10 year old children in Kazakhstan played an English word-learning game in which the robot acted as competing peer [15]. The authors investigated whether it was more effective if the robot was loosing or winning (over all, the learning was better when the robot was loosing).

Finally, the collaborative setting has been investigated in relation to fostering interactive alignment of the learning material [12,20]. The studies are different from the present one in that it employed two Nao robots in the setting with one human learner. One of the Nao robots was the teacher and the other acted as an advanced learner, and the goal was that the human learner should align his formulation of the responses to that of the robot "learner" for similar questions. The extent to which alignment occurred depended on the learner's level, with more proficient learners learning more formulations from the robot.

Compared to these previous RALL studies, the present one is not only more ambitious in terms of the intended interaction, but also in the use of the anthropomorphic robot Furhat (described in Sect. 4.3). Furhat has a human-like appearance, with realistic lip movements, which is fundamental in spoken L2 learning to convey linguistic information, and can display complex extra-linguistic human facial expressions for emotion and turn-taking signals (e.g., smile, eye and eyebrow movements). We have previously found that Furhat improved adult Swedish subjects' vocabulary learning in an unfamiliar language, Russian: the group that interacted with the robot tutor was significantly better at word retention than the groups that interacted respectively with a screen-based computer-animated tutor or with an impersonal interface that only presented the written word on screen together with audio [21].

Conversational training is more frequent in computer-assisted language learning (CALL) software, with examples such as SPELL [22], DEAL [23], The Tactical Language and Culture Training System [24] and Dansksimulatoren [25]. The latter two, which are based on the same platform, are arguably the most advanced examples of conversational training in CALL. The first allows military personnel to go through interactive skill-building lessons in Arabic combined with dialogue games using animated agents representing Iraqi citizens. The second lets L2 learners of Danish interact with a virtual community to practice spoken Danish and cultural integration. While 3D virtual environments have benefits in the large variation of tasks that can be practiced, the situated interaction with a humanoid robot incorporates other skills for face-to-face communication, such as physical rapport.

Most educational studies on L2 learning focus on human teacher-student dyadic constellations or peer collaborations in task-oriented activities. Studies on conversation, where teacher and student contributions are analyzed as those of a party, rather than as individual contributions, are still scarce. Such studies [26,27] show how both peer and teacher scaffolding are important in multiparty L2 language learning. Further, studies on peer collaborations in L2 learning settings show how important peers are for each other's learning of linguistic form and content, as well as for developing communicative skills [26,28].

Such collaboration is paramount in language cafés, but rely heavily on the scaffolding strategies of the moderator. These strategies were analyzed, through a survey of traditional language cafés (Sect. 3.1), and then through observations of how they are manifested in the simplified three-party setting corresponding to the robot-led conversation practice in this study.

3 Human Language Café Moderators

Language cafés are prolific as a resource for spoken conversational training in a second language. The variety of settings and topics during such language café settings is large, depending on the learners' L2 level, the organizer–attendee combination, and the individual moderator. The interaction ranges from quite simple questions to one participant at a time to complex whole group discussions based on a text that the participants have read.

In order to base the interaction of a humanoid robot leading a conversation session for L2 practice on the expertise of human language café moderators, we start by investigating how the latter interact with their participants and if this is influenced by factors such as experience, participant level or familiarity with participants.

3.1 Survey

A web-based survey was sent out to 140 contact persons of organizations (libraries, municipalities, NGOs, churches, universities) hosting language cafés in Sweden, encouraging them to invite their language café moderators to answer the survey. The arranging organizations and corresponding contact persons were identified through a web search for language café meetings. All hosting organizations identified through the web search that were active in arranging language cafés at the time of research and had a specified contact e-mail address were invited.

The survey contained an introductory section gathering data about the moderator and his/her participants (language practiced, gender, age of moderator and participants, organizer, language café experience of moderator and participants, participant L2 level, familiarity with participants), and then a section where the respondents were asked to describe the interaction in their language café sessions through a set of four multi-choice questions, described below. Since moderators lead sessions with different participants, the survey instructed them to choose a typical session and respond to the following questions based on this choice, if language café sessions differed.

We summarize the results of the survey, and discuss them in particular from the point of view of how strategies could be transferred to robot moderators.

3.1.1 Respondent Data

The survey resulted in 105 responses collected during a period of 3 months (November 2017–February 2018), of which all but six were submitted during the first 2 weeks. Of the respondents, 78.1% were women, 21% men and 0.9% non-binary (while exact statistics on the distribution of moderator gender is not available, data from an ongoing research project[6] suggests that women are indeed over-represented as moderators), with a large variation in age (mean age $\mu = 53$ years, standard deviation $\sigma = 16.6$, min: 20, max: 80). Most participants were judged by the moderators to be in the two age intervals 20–30 years (44.8%) or 30–40 years (44.8%), with the remainder being 40–50 years (7.6%) or 15–20 years (2.9%).

[6] Ali Reza Majlesi, personal communication,"The language café as a social venue and a space for language training".

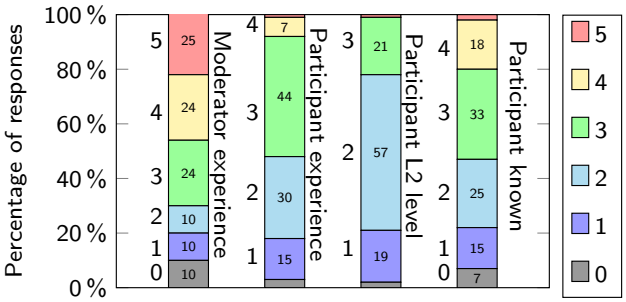

Fig. 1 Distribution of responses regarding moderator and participant experience with language cafés, the participants' level in the L2 and how well the moderator knows the participants

Almost all respondents, 100 out of 105, were hosting language cafés in Swedish, 12 in English, three each in Spanish and Finnish, two each in French, German and Arabic, and one each in a handful of other languages (some respondents were moderator in more than one language).

The language cafés were organized by libraries (35.2%), an NGO (26.7%), a municipality (9.5%) or other organizations, including churches, universities and schools (25.7%).

Figure 1 summarizes the distribution of answers concerning the moderators' and their participants' experience of language cafés, the participants' level in the practice language and the moderators' familiarity with the participants. In the following qualitative comments to the answers, we use a grouping into high (4 or 5), moderate (2 or 3) and low (0 or 1), for the different features.

On a scale from beginner (0) to very experienced (5), half of the respondents rated their moderator experience as high, one third as moderate and one fifth as low, giving a mean μ of 3.17 ($\sigma = 1.60$). The participants were rated as having less experience of language cafés ($\mu = 2.39$, $\sigma = 0.96$, on the same scale), with 70% rated to have moderate experience and almost one fifth as having little experience. Similarly, most participants were rated as having a low (25%) to moderate (80%) level of Swedish ($\mu = 2.0$, $\sigma = 0.72$, on a scale from beginners, 0, to fluent, 5).

Most (2/3) of the moderators reported that they were moderately familiar with the participants, and about one fifth each that the participants were well-known or new acquaintances every time ($\mu = 2.47$, $\sigma = 1.20$ on a scale from 0 "always new" to 5 "met many times").

In the following analysis, we attempt investigating if the moderator's or participants' language café experience, the participants L2 level or the familiarity of the participants influence the moderator strategies.

3.1.2 Moderator Strategies

The moderators were asked to estimate the proportion of a typical session that they spent using six given strategies,

determined from the authors' own experiences of language cafés. These were that (1) the moderator asks one participant at a time questions about one topic before switching to another (henceforth *Interview1*); (2) the moderator interviews several participants simultaneously about the same topic (*Interview>1*); (3) the participants ask questions and the moderator answers (*Answering*); (4) the moderator talks about a topic of her choice, such as the L2 country or herself (*Narrating*); (5) the moderator tries to encourage the participants to talk to each other (*Facilitating*); (6) the participants spontaneously talk to each other and the moderator's role is to assist when needed (*Assisting*). In addition, an Other alternative was added, for which the respondents indicated that they aimed for a normal (unstructured) social conversation; divided the group into pairs who should talk about a theme and then report back to the larger group; let the group read an easy newspaper or book and discuss what they read; gave an introduction to a topic that was then discussed in smaller or larger groups.

The moderators were asked to assign a percentage interval to each strategy (0, 1–20%, 20–40%, 40–60%, 60–80%, 80–100%). They were instructed to aim for a total of 100%, but to facilitate answering, this was not enforced. Figure 2, which summarizes the responses, may need clarifying examples: For Interview1, 33% of the moderators state that they do not use the strategy at all, 44% that they use it up to 20% of the session time, 13% use it 20–40% of the time, 6% between 40 and 60% and 2% each for 60–80% and 80–100% (summing up to 100% for the strategy). The total heights of the share bar indicate the distribution between share intervals (accumulated height of the six bars is 100%), and it can be observed that that 27% of the moderators answer that they never use some strategies (0% for in particular Answering, Interview1 and Other), that 58% of the moderators use several different strategies within the same session (the dominating share intervals are 1–40%), and that 3% use almost one strategy only (80–100% for in particular Facilitating and Other).

In order to estimate which of the strategies the moderators used the most, the number of answers for each pair of strategy and time interval, $\rho(s,t)$, was weighted with the mean of the time interval, i.e., $\alpha_t = [0, 10, 30, 50, 70, 90]$ and summarized for each strategy, $\sum_{t=0}^{90} \rho(s,t) \times \alpha(s)$, before dividing by the total accumulated time over all strategies s. To exemplify for Interview1 in Fig. 2, $\sum_{t=0}^{90} \rho(t,s) \times \alpha(t) = 33 \times 0 + 44 \times 10 + 13 \times 30 + 6 \times 60 + 2 \times 70 + 2 \times 90 = 1186$. Calculating for all strategies gives the proportions in the table in Fig. 2, indicating that the most common strategy was Facilitating followed by Interview>1, Other (i.e., social conversation, small or large group discussions on given theme) and Assisting.

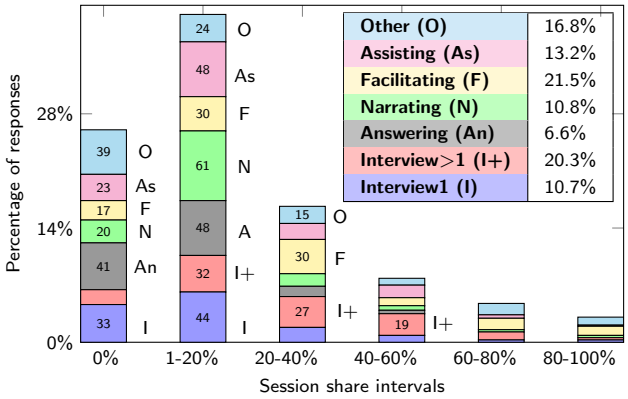

Fig. 2 Estimated time share for different moderator strategies during a typical session. Total bar heights indicate the percentage of each time interval (total accumulated bar height is 100%). Numbers within bars indicate the percentage for each strategy in that time interval (sum per strategy over all time intervals is 100%). Table: Total share for different strategies. The table also indicates the order of strategies in the stacked bars

3.1.3 Moderators' Choice of Strategy

We next analyzed if any common properties in the respondent data influenced which strategies the moderator choose. Moderator gender had little effect on the strategy, except that men spend a larger portion of the time interviewing participants, either individually or in group (35.1% vs. 30.9% for women).

For moderator age groups, the main differences were that the youngest (20–30 years) used Interview1 (10%), Interview>1 (31%) and Narrating (16%) more and Facilitating (15%), Assisting (10%) and Other (10%) less than most other age groups (the largest difference, respectively $+4, +15, +8, -5, -10, -15$%, was compared to moderators aged 41–50) and that the oldest moderators (61–70 and 71–80) used Facilitating more (27% and 22%). Participant age affected strategies minimally, with the exception of participants over 40 being interviewed less (23.9% vs. 32.5% for younger participants), with more time instead spent on Facilitating and Assisting.

The remaining factors, which are more relevant for the robot moderator experiments, are shown in Fig. 3.

Interviewing (Interview1 or Interview> 1) is used the most by the most experienced moderators, for the most experienced participants with higher level (≥ 3), and when the participants are the most familiar. The differences are however small and interviewing is a strategy that the moderators seem to deem appropriate for different participant levels and experiences.

Facilitating and Assisting were used more for participants who were more inexperienced, had a lower L2 level and were less known to the moderator. Answering and Narrating are instead used slightly more for higher level participants, prob-

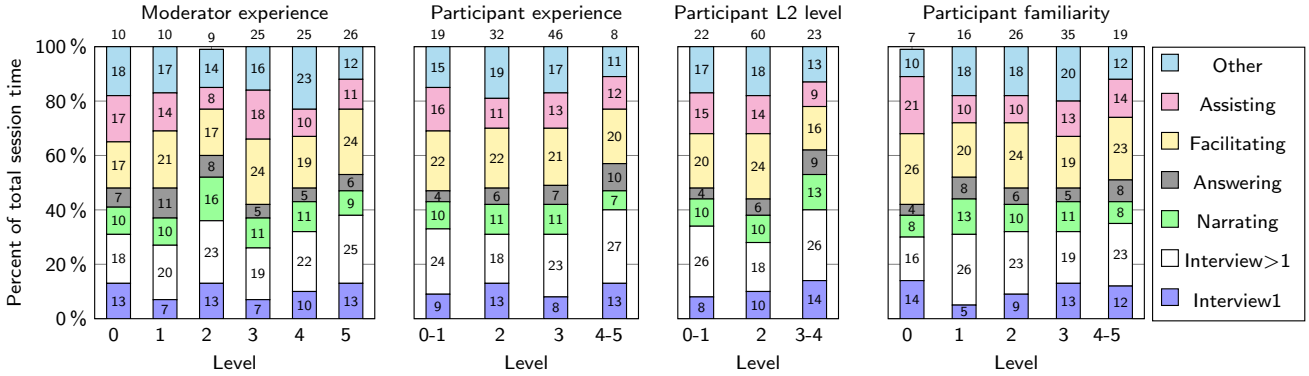

Fig. 3 Proportion of time of language café sessions spent on different strategies (legend shows vertical order in each stack), as a function of the level 0–5 of different factors. Numbers in each bar portion indicate the percentage, numbers on top of each bar how many moderators reported each level. Levels with fewer than five responses were grouped with the neighboring level

ably because the former requires more participant initiative and the latter has a more complex content.

From the survey, we hence conclude that Facilitating/Assisting and Interviewing dominate as strategies, and that their frequency of use is not directly related to the participant level. Narrating and Answering are used less, but has a role to play both in teaching the learners about the country of residence and to establish a more personal relationship between the moderator and the learners.

3.2 L2 Conversation Practice Observations

We recorded 14 short (12–15 min) and small (one native moderator and two L2 speakers) conversations on typical language café topics in a setting identical to the one that will be used for the robot-led conversations [29]. The sessions were led by 6 different moderators (A–F below; 3 female and 3 male; average age 36.2 years). Two of the moderators (B and D had previous experiences as moderators in language cafés), one (E) has been a foreign language teacher and one (A) had studied language café interaction styles, whereas two (C and F) were novices. This variation of moderator experience was introduced to observe if this influenced the interaction chosen. The instruction to the moderators was to carry out a social conversation in Swedish, and suggestions on frequently occurring topics for language cafés were provided, but the moderators were free to choose topics and interaction style.

In most of the interactions, the moderator and the two learners were unknown to each other, but the moderator knew one of the participants in sessions 4 and 13, and all three knew each other in sessions 6 and 12.

Each session was recorded with one digital video camera capturing the entire scene, and one GoPro camera and one head-mounted microphone each recording individual participants. The audio of the moderator was first transcribed using Google Cloud Speech-to-Text automatic speech recognition and then manually corrected. Each moderator turn was labeled according to which category it was considered as belonging to by the transcriber (the first author), using the labels given in Table 1.

Table 1 Interaction labels used in the transcriptions of moderator utterances in the human-led language café sessions

	Interaction label	Content
DM	Dialogue management	Social formalities and moderator-initiated switch of addressee within topic
IQ	Initial question	First question to one participant on a topic
FQ	Follow-up question	Further question to the same participant
AB	Addressing both	Open questions to both participant
RE	Responding	Longer confirmations and responses to participant utterances
NA	Narrating	The moderator tells the participants about some topic (herself, Sweden etc)
LS	Linguistic support	Short vocabulary help and longer explanations on more complicated words
BC	Back-channel	Short back-channel responses, often overlapping with participant utterance

Based on the time-aligned labels for the audio, we investigate how large proportion of the time and turns that the moderator spent on the different strategies and analyzed similarities and differences between the moderators. The time shares are shown in Fig. 4, where sessions 1–4, 5–9 and 10–11 each had one same moderator (A, B and C). The distribution of share of turns is similar to that of time, with the main differences being that back-channels, answering, follow-up

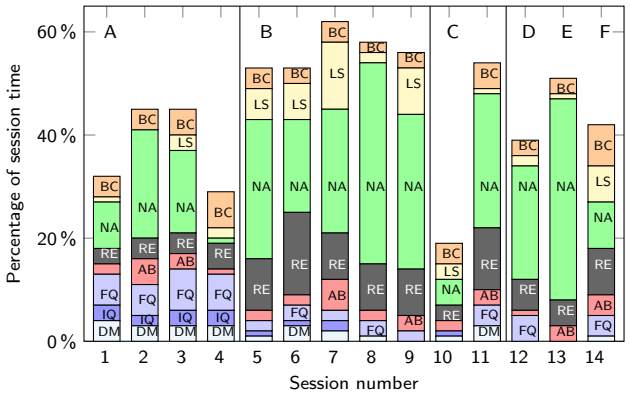

Fig. 4 Proportion of session time spent on different types of moderator utterances, defined in Table 1. Letters A–F designate moderator identity. The moderator's share of the session time is indicated by the bar height

questions and dialogue management have a larger share of turns, since they are short utterances, and narrating, which consists of longer utterances, has a smaller. The clearest prototypic examples of moderator strategies are Interviewing in sessions 1 and 3, Narrating in sessions 8 and 13, Facilitating in session 10, and three-party interaction in sessions 9 and 14.

We observe firstly that there are large differences between sessions and moderators, regarding: a) the moderators' share of session time (18–63%) and share of moderator Narrating (1–39% of session time) and b) the amount of initiative from the learners, shown by the amount of moderator Answering learner questions (3–16%). Compared to the estimated shares in Sect. 3.1, the moderators in this setting had a larger share of the session time, and spent less time on Facilitating and Assisting.

Secondly, the proportion of initial questions directed at one participant is low (1–6% of the turns). Instead, both participants were in general first addressed, but individual follow up questions were then asked (up to 19% of turns). The survey results (Sect. 3.1) that Interview>1 is more used than Interview1 are hence corroborated. On the other hand, the observations indicate a substantially larger proportion of Answering (AN) and Narrating (NA).

Thirdly, there is a large difference in the use of linguistic support (0–13% of session time) by the moderator. This is certainly linked to the participants' need for support, but moderator B clearly also used it as a strategy. He interacted with longer Narrating (note the high proportion of session time) and included more advanced vocabulary, whereupon he could ask the participants if they understood a particular word, which he then either explained, or encouraged one participant to explain to the other.

Fourthly, no evidence was found that the moderator's familiarity with the participants influenced the interaction strategy, but too little data was available for this variable to draw conclusions. There is also very little data on development of moderator strategy, but it is noteworthy that moderator C tripled her share of the session time between her first and second session, with a substantially larger part devoted to Answering and Narrating about herself, more similar to the more experienced moderators.

The importance of back-channels is also clearly illustrated by the fact that they constituted 21–54% of all moderator turns (but only 2–8% of the time).

Using the video recordings of the moderator, we further investigated how the moderator's visual attention was distributed between the two participants. The placement of the moderator and the L2 learners around the table was such that the moderator was always turned towards one participant, rather than towards both simultaneously. We therefore annotated head direction, using Anvil [30], as left or right, defining a head direction change as one when the moderator switched to look at the other participant (hence discarding head movements when the moderator was looking away to e.g., think). Note that we here strictly deal with visual attention, which is not the same as distribution of spoken turns. As will be discussed below, visual attention switches occur in different interaction situations.

Table 2 shows important differences between both sessions and moderators in how frequently visual attention was switched. This may to some extent relate to how much initiative the participants took and how verbose they were, but

Table 2 Time measurements—mean μ, standard deviation σ and maximum length—for moderator visual attention turns

(s)	Session number														Δ
	A				B					C		D	E	F	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
μ	26	15	10	21	6	5	11	6	6	10	13	8	6	10	21
σ	50	28	16	39	10	8	19	11	9	16	21	12	13	17	41
$\lvert\Delta\mu\rvert$	17	1	4	8	1	5	1	3	1	3	11	7	4	11	16
max	195	122	85	144	56	48	98	84	56	70	96	63	115	115	147
#	28	44	64	26	120	123	66	112	125	68	52	90	112	72	95

$\lvert\Delta\mu\rvert$ is the difference in the mean attention times towards the two learners, and # the number of attention shifts during the session. All sessions were cropped to the length of the shortest, 11 min 30 s, to allow for comparisons. Δ indicates the difference between the smallest and largest value in each row. A–F denote the six moderators

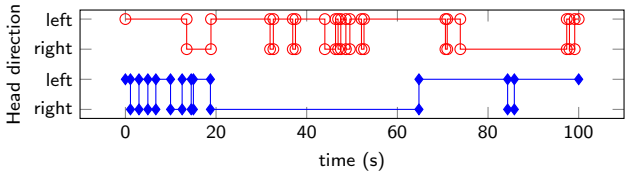

Fig. 5 Examples of moderator's visual attention: first 100 s from sessions 1 (blue, diamonds) and 6 (red, circles)

also indicates different moderator strategies. Moderator A predominantly used Interview1, whereas B to much larger extent used Interview>1 and Narrating, resulting in large differences in mean visual attention duration between the moderators (μ in session 1 vs. 5–9).

There was further a clear difference in how actively the moderators tried to host the session and balance the learners' participation. In particular moderator B very clearly tried to engage both learners simultaneously and equally, leading to very low differences in visual attention given to the two learners ($|\Delta\mu|$ in Table 2). Moderators C–F instead acted more as interlocutors and left more responsibility to the learners to take part of the conversation.

In general, the following visual attention switches were observed, sorted by duration, and illustrated in Fig. 5:

Sub-second attention: The moderator rapidly switched from one learner to the other and back again in e.g., Interview1, Answering and Assisting, to signal to the second learner that she had not been forgotten during the longer interaction with the first learner (e.g., t = 38 for session 6).

Second-long attention: The moderator consecutively switched between the two learners, signaling that both were part of the current interaction. This occurred when the moderator was Narrating, when all three were engaged in an interaction and when the moderator was following a learner–learner interaction (e.g., t = 5–20 for session 1).

Tenths of seconds long attention: The moderator used Interview1 with one learner and then turned to the second learner with similar questions. The duration depended on the length of the learner answer and if follow-up questions were asked (e.g., t = 50–100 for session 6).

Minute-long attention occurred either because the learner was very verbose when answering, or on the contrary, that the learner was slow at formulating the answer, due to linguistic problems (e.g., t = 20–65 for session 1).

3.3 Summary of Observations

The below findings on human moderator strategies are the most important when implementing a robot moderator:

The most commonly occurring moderator strategies were Facilitating (according to the survey) and Narrating (according to the observations), followed by Interview> 1 (survey and one moderator in observations) or Answering (observations), Assisting (survey) and Interview1 (survey and one moderator in observations). For the robot moderator, four strategies corresponding to, respectively, Interview1, Facilitating, Narrating and a combination of Interview>1, Narrating and Answering were implemented, as described in Sect. 4.1. Assisting is beyond current state-of-the-art capabilities and was not attempted. Answering was moreover restricted to in-topic questions that could be foreseen from the conversation context.

From the observations it appears that moderator Narrating constitutes a larger part of the conversations than the moderators stated in the survey. Different types and amount of robot Narrating were therefore tested.

There was no clear evidence that participant level, familiarity or experience of language cafés influenced the moderators' strategy. In the user test with the robot moderator, the four strategies therefore remained the same, despite differences between participant pairs, but a survey investigated if these factors influenced the participants' perception of the robot.

Finally, the observations indicated that the human moderators used visual attention switches of different duration, depending on the conversation state. These measures of moderator attention were not directly implemented in the robot's strategies, but will be used for comparison in Sect. 5.2, when analyzing how the robot's interaction style influenced its attention behaviour to address a single participant or both of them.

4 Robot Language Café Moderators

Having performed this analysis of human language café moderators, we implemented a set of four corresponding strategies and conducted a user study of a robot moderator interacting with pairs of L2 learners of Swedish.

We first describe these different strategies (Sect. 4.1) before giving details on the implementation (Sect. 4.2), the participants (Sect. 5.1) and the user experiment conducted (Sect. 5).

4.1 Interaction Strategies

The four strategies shown schematically in Fig. 6 and exemplified in Table 3 were implemented in the social robot Furhat (c.f. Sect. 4.3). Three of the four settings are the cardinal points of the space spanned by the dialogue dimensions Initiative and Focus, i.e., if the robot or the learners lead the interaction and if the topics of conversation focus on the learners or the robot. The fourth cardinal point, robot focused interaction with learner initiative (corresponding to Answering in Sect. 3) was replaced by Interlocutor, as it was deemed

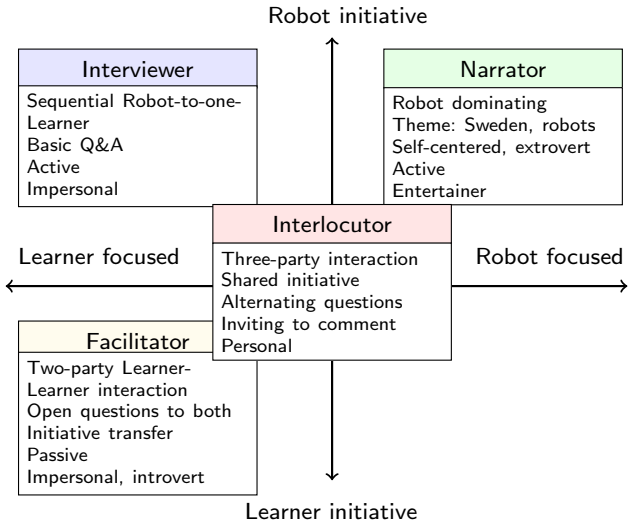

Fig. 6 Schematic overview of the four robot moderator strategies

that it was not possible to ensure that the robot could successfully answer unrestricted learner questions.

It should be noted that the goal is not to create the best possible interaction during a session, but to investigate how the learners perceived the different strategies separately. Each strategy was therefore made as distinct as possible and maintained during one entire session.

4.1.1 Interviewer

The Interviewer strategy primarily consists of addressing one participant at the time, with short, direct questions, e.g., *"Which languages do you speak?", "What are your hobbies?, "What did you do this weekend?"* etc. After having asked the same learner a number of connected questions, the robot turns to the other participant (hence corresponding to Interview1 in Sect. 3). The robot has the initiative and drives the conversation focused on the learners, with the robot asking questions without providing much information about itself or its own opinions. If asked such questions by the learners, it will in fact attempt to not answer.

The underlying pedagogical concept is that asking one participant one simple question creates a clearly structured and comfortable practice environment, in which turn-taking is defined by the robot and the expected content frame for each answer is set. The setting was created to primarily be suitable for basic level learners, but as demonstrated by the human moderator strategies, interviewing can in fact be of use also for learners at a higher L2 level.

4.1.2 Narrator

In the Narrator strategy, the robot's goal is to convey either its opinions (e.g., about robots' role in society in the future) or some knowledge (e.g., about Sweden) to the learners. Furhat will ask for feedback on its opinions or for answers to social or trivia questions, but its focus is to continue with the own narrative, regardless of the input from the learners. In this setting, the robot has an egocentric extrovert entertainer personality and provides much more information about itself and makes jokes. The robot maintains the initiative, and the dialogue is unbalanced with the robot talking most of the time. The content of the session may range from a semi-monologue to a collaborative quiz game, where learners discuss answers to the robot's questions.

The underlying didactic concepts for this setting are firstly listening comprehension, secondly that transfer of realia may be motivating and thirdly that the learners' engagement in interacting with the robot may increase if the robot is having more personality.

4.1.3 Facilitator

The Facilitator strategy is the opposite to the Narrator in focus, initiative and personality. As Facilitator, the robot's goal is to get the two learners to talk with each other, and to interfere as little as possible, unless required to stimulate the dialogue (e.g., asking one learner to comment on what the other said or suggesting topics to continue discussing). The robot will ask more general, open questions, addressed at both learners simultaneously and will also encourage the learners to choose topics to discuss. This signifies that the robot should be much more passive than in the two previously described settings and is also more introvert and impersonal.

The underlying pedagogical idea for Facilitator is that the learners, rather than the robot, should be as active as possible during the practice. It should be noted that the robot facilitator is only facilitating the dialogue flow, not providing any linguistic assistance. This setting requires that learners take more initiative in the dialogue and that both are able to carry out a conversation with each other.

4.1.4 Interlocutor

The Interlocutor setting is the one that corresponds the closest to human moderator behaviour in Sect. 3, as the robot tries to establish a three-party interaction, with alternating questions to both learners, inviting them to comment on each others statements and providing own input. The initiative is more shared, with more open questions to both learners (Interview>1, Addressing Both in Sect. 3). The robot is also more personal, both in terms of providing information about itself, and in being more personal when addressing the learners (e.g., by calling them and their countries by name, as exemplified in Table 3).

The underlying pedagogical concept is that the similarity with human interaction, where all three parties are included and personal relationships are formed, should be beneficial

Table 3 Stylized example of differences in dialogue flow for the different robot strategies, with two fictional learners, A and B, and the robot (R)

Turn	Interviewer	Narrator	Facilitator	Interlocutor
1	**R:** [→A] **What country are you from?**	**R: Do you know how many inhabitants Sweden has per square kilometer?**	**R: How different are your home countries?**	**R: A, are there differences between Sweden and France?**
2	A: *France.*	A: *I do not know.*	A: *Where are you from?*	A: *People are more open in France.*
3	**R: How long have you been in Sweden?**	**R:** [→B] **Can you guess?**	B: *I am from Syria. And you?*	**R:** [→B] **B, do you find it difficult to get to know Swedes?**
4	A: *Almost 2 years.*	B: *One hundred?*	A: *I am from France. It is not as warm as Syria.*	B: *Not really, I found many Swedish friends almost directly.*
5	**R: And what do you do in Sweden?**	**R: No, much less, only about 20. And most of the inhabitants live in the cities in southern Sweden.** [→A] **Have you been somewhere else in Sweden?**	B: *And there is no war in France.*	**R: Did you do anything special to get to know people, B?**
6	A: *I study Chemistry at the university.*	A: *I have been in Visby twice.*	A: *No, that is right. Did you come to Sweden because of the war?*	B: *Social events at the library are quite good to get to know locals.*
7	**R: Interesting!** [→**B**] **Do you also study?**	**R: I have also been there! I imitated a Swedish minister at a political fair in Visby. I think I might have a political career. Do you think that it would be good to have robot politicians?**	B: *Yes, we had to flee.*	**R: That can be a good way.** [→**A**] **Do you also do this, A?**
8	B: No, I work at a supermarket.	B: *Why should we have that?*	**R*: How awful. Do you want to tell us more?**	A: *No, there is no library close to where I live.*
9	**R: How long have you worked there?**	**R:** ↔ **To fight for robot rights. Do you think that robots should have rights?**	B: *We walked and walked for many nights.*	**R:Do you have any other tips on how to get to know people?**
10	B: *I have worked there since January.*	A: *No! They have no real feelings and senses.*	A: *How did you manage to get to Europe?*	B: *I started playing football.*
11	**R: And how long have you studied Swedish?**	**R: Now you are really hurting my feelings!**	B: *We were lucky and found a good boat.*	**R:** [↔] **I also like football. But I can't play. I don't have any legs!**

The R* in turn 8 in the Facilitator setting indicates that the robot may take the turn, but only if A does not take it. → indicates that the robot turns it head towards one of the learners, ↔, in turn 9 of Narrator setting, that it is addressing both

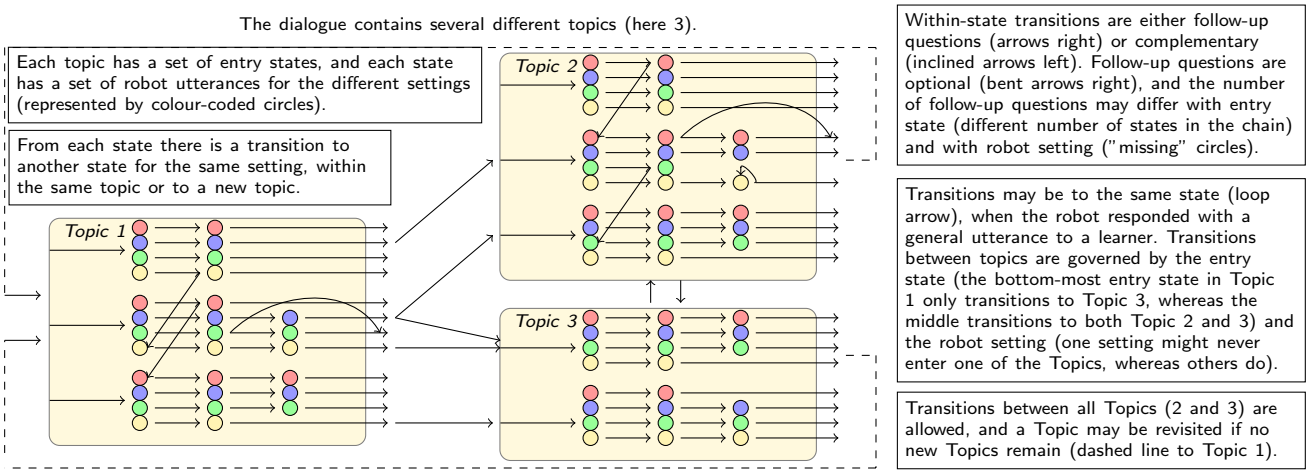

Fig. 7 Schematic overview of the dialogue flow in the language café sessions. The flow is similar for the four settings

for practice and motivation. Compared to Interviewer and Narrator, more learner initiative is required.

4.2 Implementation

For each turn in the dialogue, the robot settings had a personality-specific set of utterances and possible transitions to the next set of utterances, as illustrated schematically and explained in Fig. 7.

All utterances were pre-generated for each state, but the set-up with topics, states and optional transitions nevertheless allows for large flexibility in the dialogue. Moreover, as most state-setting combinations have alternative utterances and utterances that had already been used were pruned, as the robot has a repertoire of general response utterances (*"Yes", "No", "Mm", "Mhm", "I do not know"*) allowing it to react to learner questions and comments, and as the utterances are personalized to include learner names, countries and languages in the Interlocutor setting, the conversations could be quite different. This is a general prerequisite for maintaining learner interest, and of particular importance for our user test, where each subject should experience each setting. We want to avoid repeating the same dialogue flow for the same learner, since this would lead to a bias, as learners would feel that interaction settings coming later were repeating previous sessions.

For the user experiments described in Sect. 5, a semi-automated wizard-of-Oz set-up was used, in which a human controller listened to the conversation and selected one of the robot utterances. This set-up was used to prevent that technological problems, such as failed ASR of the participants' utterances, influenced the study that should focus on differences that are due to the robot's interaction style. The robot was controlled using an interface (c.f. Fig. 8) specially developed for this experiment. At each dialogue transition, a set of utterances was loaded into the interface and the wizard used keyboard short-cut keys to choose robot utterance or to turn the robot's head (in most cases, head turns were automated to accompany utterances, as exemplified by turns 3, 7 and 11 for Interlocutor in Table 3). The interface further included a short-cut key to repeat the previous robot utterance (with maintained speaking rate and emphasis) and a web-cam live-stream of the whole scene (c.f. Fig. 8), to allow the wizard to monitor non-verbal signals from the participants.

The choice of only presenting a limited set of 10 utterances to the wizard was based on our previous study [29], where long robot response times were seen as the major problem. They were caused by the wizard choosing from a larger set of utterances from any topic in a more complex interface, or generating a custom robot utterance, using speech recognition or by typing, followed by speech synthesis for the robot's output. Even if the possibility of typing in an answer was available in the present interface the wizard was discouraged from using it in order to keep the interaction pace (and the possibility was in fact never used in the user study below). In states for which more than 10 utterances, including transitions to new Topics, were available, 10 utterances were randomly selected. The wizard was aware of which setting the robot had for each session and aimed at maintaining the best possible conversation with the available utterances, while maintaining the distinctive features of the setting.

The implementation was done using FARMI [31], combined with a python-wrapper to IrisTK [32][7] allowing access to components for text-to-speech synthesis (a Cereproc TTS voice was used), facial animation and interaction event tracking. Since the data recording is time-synchronized it is possible to replay the streams recorded (audio, video and

[7] https://github.com/jonepatr/furhat-client.git.

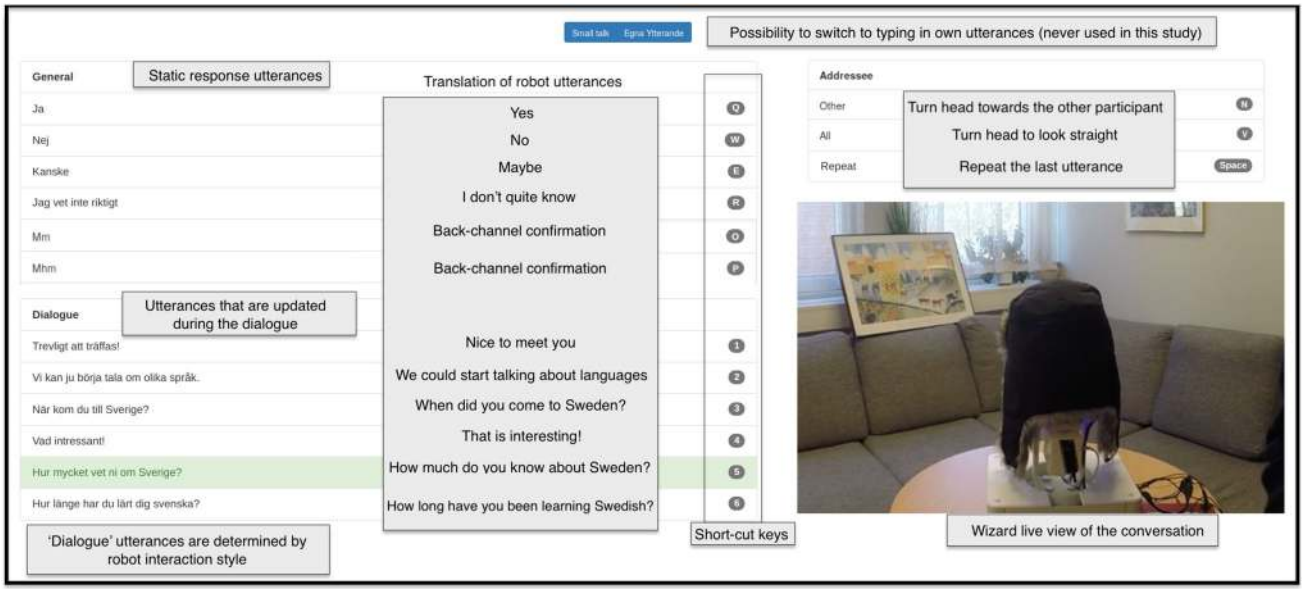

Fig. 8 The semi-automated wizard interface. Grey-boxed text provides translations and explanations

action selection) off-line, which may be useful in the future to automate robot functions.

4.3 Furhat

The Furhat robotic head [33], can display a wide variety of human face gestures, as it uses a computer-animated face projected on a 3D mask (c.f. Fig. 9). As the neck is fitted with a motor-servo, Furhat can also turn its head, which has been shown to be important in three-party interactions [34] and collaborative or competitive quiz games [35]. Other studies with Furhat have e.g., focused on using the robot for interaction with the elderly to detect early signs of dementia [36] or act as a simulated Alzheimer patient to investigate interlocutor responses [37].

5 User Experiment

A within-participant study was performed in an office at the premises of one provider of SFI courses in Stockholm.

The experiment set-up was that two L2 learners of Swedish were seated next to each other at one side of a round café table and Furhat placed on the table opposite to the learners, as shown in Fig. 9. Head-mounted microphones recorded the audio and web-cams a video of each participant. The audio-visual recordings will be used for future interaction analysis, for ASR adaptation and for training of motivational state detection. All robot utterances were also logged for further analysis.

The wizard-of-Oz (the first author) was seated at a desk placed behind an office room divider next to be café table,

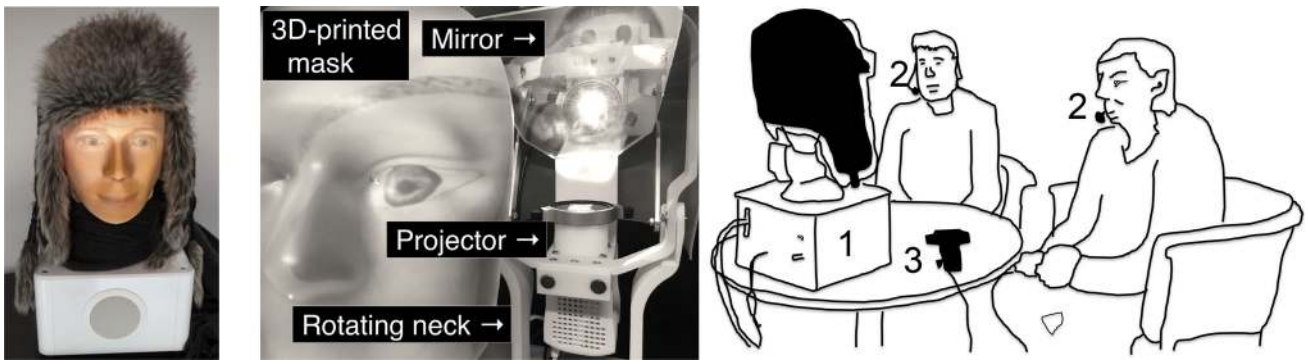

Fig. 9 The Furhat robot (left), with face removed to show the projection components (middle) and a stylized artistic drawing of the experimental setup (right). The drawing indicates: (1) The placement of the Furhat robot on the table, (2) The head-mounted microphones for each of the subjects, (3) one of the web-cameras, capturing the left subject. An additional camera placed on the table (hidden by Furhat in this view) captures the right subject, and another one, placed behind Furhat captures the whole scene (corresponding to the present view)

controlling the verbal and non-verbal behavior of the robot, using the wizard interface shown in Fig. 8. The participants were told that the first two authors, who were present during the conversations, were controlling the audio and video recordings, and all subjects believed that they had interacted with a fully autonomous robot.

The experiment with each user was divided into two sets, one on day one (with two sessions of 10–15 min each with two different robot interaction settings) in one learner pair, and one on day two (with the remaining two robot settings) with another peer learner.

5.1 Participants

The user study was performed during 3 days in March 2018, in conjunction with the learners' SFI classes.

The participants had been informed by their teachers about the user study and learners judged to be at least at B1/B2 level in the Common European Framework of Reference for Languages (roughly corresponding to level 2–3 in Sects. 3.1 and 5.3) had been invited to sign up for one 30 min session with the robot and one other learner on day 1, which automatically assigned them to a second session on day 2, with a different learner. 32 subjects were thus recruited, of which four were discarded in an initial screening test because their level of Swedish was too low for them to be able to interact in spoken conversations with a peer and the robot for 10 min (of the remaining 28 subjects, several were at level A2).

All subjects were introduced to the conversation session by one of the authors telling them that the experiment was intended for the robot to learn how to behave in different types of dialogues. The participants were not given any information about the different robot settings. They were informed about how the experimental data would be handled and signed an informed consent form.

Of the 28 subjects, 20 came to both sessions that they had signed up for (attendance is not compulsory at SFI and the subjects who dropped out of the study did so because they could not come to class on day 2, not because they actively opted out of the study). To compensate for the fact that 8 of the subjects thus had participated in sessions with the first two settings only and to allow the remaining 20 subjects to complete all four conversations, an additional 4 subjects were recruited. In addition, one subject only had time to participate in one of the conversations in set 2 and was therefore replaced in the last session by the second author (who is a proficient L2 speaker of Swedish).

This means that the study in total included 33 participants (18 female, 15 male, mean age 32 years, $\sigma = 8.6$ years), of which 19 experienced all four settings, 1 three settings and 12 two settings (the second author is not included in the analysis) and 48 recorded sessions.

The number and order of robot settings were initially balanced, but due to the subject drop-out and the need to recruit new subjects, some imbalance was introduced and in total 23 Interviewer (as first setting: 9, as second: 6, as third: 5, as fourth: 3), 27 Narrator (1st: 10, 2nd: 8, 3rd: 8, 4th: 4), 27 Facilitator (1st: 9, 2nd: 8, 3rd: 6 and 4th: 5) and 26 Interlocutor (1st: 4, 2nd: 10, 3rd: 2, 4th: 10) learner experiences were rated.

The participants were from Syria (5), two each from Iran, Iraq, Egypt, Afghanistan, Ukraine, Albania and Poland, and one each from Azerbaijan, Chile, China, Congo, Croatia, Cuba, Eritrea, Italy, Kazakhstan, Kurdistan, Philippines, Somalia, and Spain. The self-reported first languages were Arabic (10); Spanish (3); Ukrainian, Russian, Polish, Italian (2 each); and one each of Chinese, Croatian, Dari, Filipino, French, Greek, Kurdish, Persian, Portuguese, Punjabi, Somali and Tigrin.

5.2 Dialogue Data

In order to allow for comparisons with the human moderators' attention shifts towards the learners, the time intervals between head turns for different robot settings were determined. Contrary to the experiment in Sect. 3.2, due to the limited physical space used, the placement of robot and learners was such that the robot could address both learners by looking straight ahead. We observe in Table 4 that the differences in mean and maximum attention times were substantial and linked to the interaction of each strategy: As Interviewer and Interlocutor the robot addressed one single learner at the time much more than both, but as Narrator it almost exclusively addressed both learners, and as Facilitator both to a large extent. As intended, Interlocutor had more frequent attention switches than Interviewer.

Compared to the human moderators in Sect. 3.2, we see that the mean attention times towards one learner were long for Interviewer and Facilitator ($\mu = 59, 58$s, compared to 26s for the longest human moderator mean) and comparable for Interlocutor and Narrator ($\mu = 23, 17$s). Even if the human moderator attention shifts were visual and did not necessarily correspond to a shift of verbal attention, we note that the duration of the robot's attention towards one single learner may be perceived as unnatural and may need to be shortened ($\mu_{hum} = 11$s). We further observe that Narrator had very few head turns, and could potentially benefit from more frequent visual attention shifts to connect with the interlocutors.

5.3 Survey

After each conversation, the participants used a tablet to fill in a short web-based questionnaire about the dialogue. The questions included how they would describe

Table 4 Robot attention duration—mean μ, standard deviation σ and maximum—towards a single learner, or both of them

(s)	Robot setting							
	Interviewer		Narrator		Facilitator		Interlocutor	
	Single	Both	Single	Both	Single	Both	Single	Both
μ	**59**	17	*17*	**365**	58	157	23	*8*
σ	38	30	*15*	**201**	**93**	143	27	*19*
$\Delta\mu$	*-42*		**186**		90		−8	
Max	192	130	*41*	**626**	437	544	**238**	*110*
μ#	11		*1.9*		6.8		**27**	

$\Delta\mu$ is the difference in mean attention times between turns directed at both learners compared to towards one single, and μ# the average number of moderator head turns during a session. The smallest (italic font) and largest (bold font) values for single and both are highlighted

· the distribution of the content of the session: *"the robot asked one learner questions at a time"* (*Interview1*), *"Learner-learner conversation"* (*Facilitating*), *"The robot talked about a topic of his choice"* (*Narrating*), *"Conversation between all three (learner, learner, robot)"* (*Participating*).
· who had the initiative
(*"Always the robot"*, 0—*"Always the human learners"*, 5)
· how friendly the robot was
(*"Unfriendly"*, 0—*"Very friendly"*, 5)
· how personal the robot was
(*"Keeping a distance"*, 0—*"Too personal"*, 5)
· the robot's interaction behavior
(*"Extremely machine-like"*, 0—*"As a human"*, 5)
and how they would rate
· the robot as a conversational partner
(*"Extremely poor"*, 0—*"Excellent"*, 5).
· the session from a learning perspective
(*"Poor"*, 0—*"Excellent"*, 5)

Data was also collected on the participants language café experience (*"Never"*, 0—*"Often"*, 5: $\mu = 1.25$, $\sigma = 1.9$), self-reported level of Swedish (*"Beginner"*, 0—*"Fluent*, "5", $\mu = 2.1$, $\sigma = 1.1$), and if the peer in the conversation was familiar (*"Never met"*, 0—*"Close friends"*, 5).

5.4 Survey Results

The participants' perception of the content of the sessions was first investigated. As for the human moderator survey, the respondents should indicate a time proportion (0%, 1–20%, 20–40%, 40–60%, 60–80%, 80–100%) for each of the four descriptions and were instructed that the sum should be approximately 100%. As many answers did not sum to 100%, the time proportions for the four categories were normalized for each subject–session combination. The resulting proportions, shown in Fig. 10, were similar for the different robot settings, and a χ^2-test showed that there was no significant

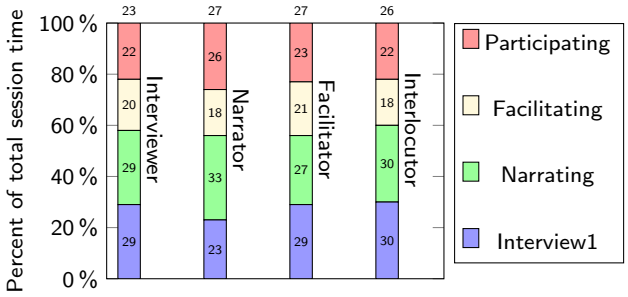

Fig. 10 Perceived proportion of time of the conversation that the robot spent on different interaction styles, according to the participants, for the four robot setting. Numbers in the bar portion indicate the percentage, numbers on top of each bar indicate the number of subjects for each setting. Legend indicates order in each stack

difference between the settings. The participants could hence not tell the content of the interaction strategies apart.

This is less surprising than it may seem at first, for several reasons:

Firstly, the robot interaction styles are not disjunct. All conversations included elements from several interaction styles, e.g., the robot presenting itself (Narrating) and asking for learner input (Interview1 or Participating). For example, Interlocutor asked questions also to individual learners (Interview1) and provided information about itself (Narrating), Narrator asked for answers to quiz questions (Interview1), and Facilitator for suggestions on topics for discussion (Participating).

Secondly, labels may not have been self-evident for the respondents, e.g., as Interviewer, the robot asked questions *"about a topic of his choice"*, which could fit the Narrating label, and it did distribute the questions to both learners approximately evenly during the session, which respondents may have considered to be a *"Conversation between all three"*.

Thirdly, the robot's interaction differed between sessions, even within the same interaction style, e.g., depending on if the Narrator focused on quiz questions, an egocentric dialogue or a monologue.

Fourthly, depending on responses (or lack thereof) from the learner pair, the wizard sometimes had to fall back on a less distinct interaction style, e.g., taking more initiative to lead the conversation in the Facilitator setting, if the learners did not engage in learner–learner conversation.

It can nevertheless be observed that the Narrator setting was perceived as containing less Interviewing (22% vs. 29–31%) and more Narrating (34% vs. 26–30%) compared to other settings.

Next, survey responses were analyzed with respect to between robot setting differences. A single factor ANOVA for robot setting over all subjects ("all" below) showed no significant differences between robot settings, neither when the survey categories (Initiative, Friendliness, Personal, Human-like behaviour, Conversational behaviour, Learning value) were considered separately, nor when they were clustered by robot setting, as an overall quality score given by each subject. For the clustering, the ratings for Initiative were excluded, since they constitute the subjects' description of the interaction, rather than preference scores. However, the corresponding single factor ANOVA for the 19 subjects who experienced all four settings, ("completion group" below), showed a significant effect of robot interaction strategy for the clustered ratings. A Tukey post-hoc analysis revealed that Interviewer was rated significantly higher than Narrator and Facilitator. Moreover, a two factor ANOVA for the completion group indicated significant differences between both robot strategies ($p < 0.05$) and subjects ($p < 0.01$) and their combination ($p < 0.001$). Even if there was no significant difference in the subjects' perception of the *content* of the sessions with different strategies (c.f. Fig. 10), there was a significant difference between their perception of the *quality* of different settings. We therefore explore the underlying dependencies for these differences. Figure 11 summarizes the ratings in terms of means and standard deviations of the ratings and the main observations can be summarized as (if not otherwise stated, the observations hold for both the "completion group" and "all"):

Interviewer was the setting with the highest mean for Learning. It was further perceived as the most Friendly and the most Personal. For these there factors, it also had the lowest standard deviation in the ratings. It further received the highest rating for Conversing behavior, even if the learners identified that it was the setting in which the robot had the initiative the most (over all answers) or the second most (over the completion group). It should be noted that one contributing factor to the preference for Interviewer could have been the setting in conjunction with SFI classes, as this may have induced the learners to expect a more classroom-like interaction.

Narrator received the next highest ratings from a learning perspective (together with Interlocutor for the completion group) and regarding how Personal the robot was. It was on the other hand perceived as the least Friendly, and as being the one that held the Initiative the most (completion group) or second-most (all). The large standard deviations in the answers are further discussed in Sect. 5.5.

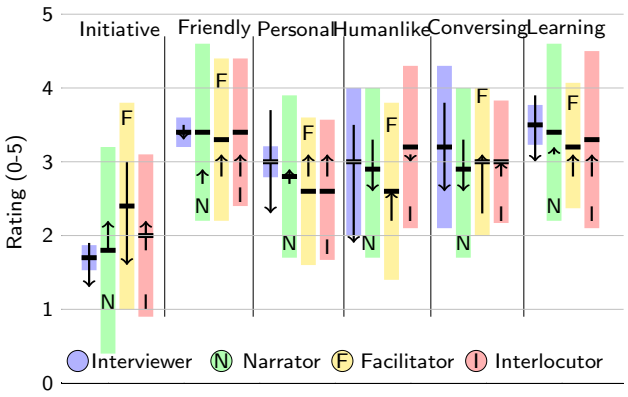

Fig. 11 Participant responses for the six survey questions. Horizontal bars show the mean, vertical coloured bars the standard deviation around the mean. The arrows indicate, for the 19 respondents who experienced all four settings, the difference between respondents who experienced a setting in one of the two initial conversations (start of arrow) and those who experienced it in the two final (end of arrow). None of the within-category differences are significant. (Color figure online)

Facilitator was identified as the setting where the learners had the most Initiative, but was rated lowest for Learning and regarding how Human-like the robot's behavior was. This was most probably because the robot could sometimes not respond to questions or suggestions by the learners in the less controlled dialogue. Nor could it provide the Assisting support that the human moderators provided when acting as facilitators.

Interlocutor was rated the highest in Human-likeness and Friendliness, but lower for Learning, and, somewhat surprising, regarding how Personal the robot was.

As illustrated by the arrows in Fig. 11, there are substantial differences between responses for several of the settings depending on if it was experienced as one of the first two settings or as one of the last. It should be noted, however, that these comparisons are between different respondents and that there is a general decrease in rating, except for Initiative, when comparing conversations 1–2 with 3–4 (Learning: -0.3, Friendliness: -0.1; Personal: -0.3; Conversation skill: -0.2, Human-likeness: -0.6).

The reasons for this decrease may be that the first enthusiasm of interacting with a robot fades (this was found, for a longer time frame by [16]), but it is more probable that sessions 3–4 to some extent felt like repeating conversations 1–2. Even if there were many different paths to be taken in the dialogues and the robot utterances and their transitions differed between the settings, most of the topics and the general flow of the dialogue were nevertheless common for all settings. It is hence natural that a participant who gets the same

Table 5 Respondent ranking order for the different robot settings and session number for the different dimensions in the survey

	Ranking of robot interaction settings				Session
	1st	2nd	3rd	4th	
Learning	Interviewer	Interlocutor	Facilitator	Narrator	2 1 4 3
Friendly	Interviewer	Facilitator	Interlocutor	Narrator	2 3 1 4
Personal	Interviewer	Narrator	Facilitator	Interlocutor	2 3 1 4
Initiative	Facilitator	Interlocutor	Interview	Narrator	2 4 1 3
Conversing	Interviewer	Interlocutor	Narrator	Facilitator	2 1 4 3
Human-like	Interviewer	Interlocutor	Facilitator	Narrator	2 3 1 4

types of questions a second time, in particular by the same robot, will rate the second session lower (one participant in fact told the robot *"But I already told you this yesterday"*).

Despite this caveat, one may observe some trends in the differences in answers, not the least that Interviewer is rated lower for Learning, Personal, Human-likeness, and Conversing behavior, after having experienced more of the other settings. The trend is the opposite for Facilitator, while Interlocutor show similar trends as Interviewer.

In order to investigate within respondent preference, the rank sum was calculated for each dimension and each respondent separately. The highest respondent rating for each dimension was assigned 1 and the lowest 4, accounting for ties; e.g., a set of user ratings of (4, 5, 4, 3) gives the rank score (2, 1, 2, 4). These rank scores were then summed by robot setting and by dialogue number, giving the ranking orders shown in Table 5. Interviewer was ranked the highest for all dimensions but Initiative. For Narrator, the results differ for Learning, as it received the next highest ratings (Fig. 11), but was ranked the lowest, indicating that most learners preferred the other settings, but those that did prefer Narrator rated it highly. We will return to this issue below (in particular in Sects. 5.5, 5.10 and 5.12). An artifact of dialogue order may also be observed, with dialogue 2 getting the highest ranking along all dimensions.

5.5 Influence of Perceived Interaction Style

As discussed above, from the learner's perspective, the differences in the robot's behavior between the four settings was not distinct, and they rather experienced varying degrees of different content of the learning session (Fig. 10), which influenced their rating (Fig. 11).

We therefore analyzed how the subjects' perception of the robot's behavior (measured as the share of session time for the four behaviors Interview1, Narrating, Facilitating and Participating) during the session correlated with their ratings. The correlation was calculated over the 103 subject responses (19 subjects $\times$ 4sessions + 1 subject $\times$ 3 sessions + 12 subjects $\times$ 2 sessions) and a Pearson's correlation test was used to test significance. This method was used, rather than a repeated measures ANOVA since the data for the variable robot behaviour was complete for only 19 of the 32 subjects. Table 6 shows that there are a number of significant correlations between the session share of each robot behavior and the ratings:

A positive correlation between the amount of Interview1 and the ratings on Conversing behavior ($p < 0.01$), Personal and Human-likeness ($p < 0.05$).

A negative correlation between the amount of Narrating and the ratings on Learning, Personal and Human-likeness ($p < 0.01$), and Conversing behavior ($p < 0.05$).

A positive correlation between the amount of Narrating, Facilitating and Participating and the perceived amount of user Initiative ($p < 0.05$).

The large standard deviation for Narrator in Fig. 11 and the significant negative correlation between Narrating and rating in Table 6 need to be discussed further.

We analyzed the robot utterances during all the Narrator sessions to investigate if there was any connection between the actual robot utterances, the perceived share of Narrating and the rating for Learning. The per session ratings and corresponding perceived amount of Narrating (both averaged over the two participants) were one 5 (30% Narrating), one 4 (25% Narrating), five 3.5 (30% Narrating), six 3 (38% Narrating) and one 2.5 (30% Narrating). The highest rated Narrator session can be said to have been quite close to an Interlocutor behavior: the robot used several different interaction strategies within Narrator (asking trivia questions, narrating about himself, asking the participants for their views, but then responding with his views or a topic shift, rather than asking follow-up questions), leading to a quite social conversation, albeit with an egocentric interlocutor. The lowest rated Narrator session mainly consisted of two components: Furhat talking about itself and robots; and trivia questions, but this was not particularly different from higher rated sessions.

However, when analyzing the average robot turn length (here measured as the average number of characters per utterance), a possible explanation appears, as the two highest rated Narrator sessions had the lowest average (43.1 and 47.0 characters), while the one rated 2.5 had among the highest (61.8, maximum 68.2). As a comparison, the average turn length in

Table 6 Correlation between the session share of different robot behaviors, as perceived by the subject, and the rating of the session

	Interview1	Narrating	Facilitating	Participating
Learning	0.17	−0.34**	0.18	0.06
Friendly	0.09	−0.10	0	0.03
Personal	0.24*	−0.33*	0.06	0.07
Initiative	−0.03	0.24*	0.24*	0.24*
Conversing	0.28**	−0.23*	0.10	−0.14
Human-like	0.21*	−0.32**	0.14	0.07

Level of significance: $^{*}p < 0.05$; $^{**}p < 0.01$

Interviewer sessions was 34.2 characters. We hence deduce that a main reason for the lower scores for Narrator may be the length of the Narrator utterances, and the thus induced difficulty. The long Narrator utterances may be problematic from both a pedagogical and a technological perspective. The pedagogical problem could be that the learners disliked the more passive role in this setting; the technological that the longer robot utterances were more difficult to understand when generated with TTS. This hypothesis is supported by the free text answers, in which Narrator received as many comments (4 out of 8) as the other three settings together stating that the robot talked too fast.

5.6 Influence of Participant Variables

We have further used factorial analysis to analyze if any factors related to the participants, such as L2 proficiency, familiarity with language cafés, age range, gender, first language and if they knew the other participant, influenced their ratings of the session. Since a repeated several factor analysis could not be performed, as not all of the subjects are equally represented in the data (as there are only 2–3 answers from 13 of the subjects), separate one-factor analyses are made. It should be noted that the number of data points (survey answers, n below) per category becomes low in several cases and the standard deviation is large (due to large variation in the responses). We are therefore mostly only able to observe trends, rather than finding significant differences between different categories. These trends are summarized below.

5.7 L2 Level

The ANOVA showed no significant differences, but the trend is that learners at levels 1 and 2 rated Learning higher ($\mu_1 = 3.7, \mu_2 = 3.5, \sigma = 1.0; 1.0, n = 14; 26$) than those at lower and higher levels ($\mu_0 = 3.3, \mu_3 = 3.1, \sigma = 1.6; 1.0, n = 10; 53$). This is in line with the findings in [29], where we found that participants need to have a basic level in order to benefit from the conversation practice, but that more proficient learners require a more advanced conversation than is currently provided by the robot moderator.

For the combination of robot setting and participant L2 level, the following (non-significant) trends were observed for Learning:

Interviewer had a peak at level 2 ($\mu_2 = 3.8, \sigma = 1.3, n = 5$), compared to level 1 and 3 ($\mu_1 = 3.3, \mu_3 = 3.4, \sigma_{1,3} = 1.3, 1.1, n_{1,3} = 4, 13$).

Narrator had a clear negative trend ($\mu_0 = 4.0 \rightarrow \mu_3 = 2.9, \sigma = 0; 0.64, n = 3; 13$), potentially indicating that more proficient learners want to contribute more actively (as they can be assumed to have understood the TTS better than lower level learners, the problem is probably not the difficulty of the utterances).

Facilitator had a higher rating for level 2 than for other levels ($\mu_2 = 3.6, \sigma = 0.90, n = 8$ vs. e.g., $\mu_3 = 2.8, \sigma = 0.80, n = 12$). We tentatively attribute this rise and drop to the fact that the lowest-level participants were not proficient enough to take more initiative in the dialogue, which therefore became halting, whereas the more proficient ones did take more initiative, but rated the setting more negatively if the robot did not reply to their questions.

Interlocutor had a very large variation in the ratings. It was rated the lowest of all settings by participants at levels 0 and 2 ($\mu_0 = 2.5, \sigma = 2.1, n = 2$; $\mu_2 = 2.8, \sigma = 0.80, n = 5$), but the highest by those at level 1 ($\mu_1 = 4.7, \sigma = 0.60, n = 3$) and second by those at level 3 ($\mu_3 = 3.2, \sigma = 1.1, n = 15$).

5.8 Age Group

The responses were pooled into the three age groups <30 years (n = 39), 30–40 years (n = 42) and >40 years (n = 22). The ANOVA showed no significant effects for age groups, but the rating of Learning increased with age, $\mu_{<30} = 3.2, \mu_{30-40} = 3.3, \mu_{>40} = 3.5$ ($\sigma = 1.1; 1.1; 0.91$).

The youngest group preferred the Interlocutor setting ($\mu_{<30} = 3.4, \sigma = 1.4, n = 10$), while the two other groups preferred Interviewer ($\mu_{30-40} = 3.6, \sigma = 1.0, n = 11, \mu_{>40} = 4.0, \sigma = 1.2, n = 4$). Facilitator ($\mu_{<30} = 2.9, \sigma = 1.0, n = 10$) and Interlocutor ($\mu_{30-40} = 3.1, \sigma = 1.1, n = 11$; $\mu_{>40} = 3.3, \sigma = 1.0, n = 4$) were rated lowest in the respective groups.

5.9 Language Café Experience

The distribution of previous experience of language cafés was rather uneven ($n_0 = 64$, $n_1 = 4$, $n_2 = 10$, $n_3 = 9$, $n_4 = 4$, $n_5 = 12$), and to increase the number of data points per category, the participants were grouped into the three groups "no experience" (0), "some experience" (1,2,3) and "much experience" 4,5).

With this pooling, we observe that participants with no or little experience of language cafés have a similar view of the Learning effectiveness of the practice ($\mu_0 = 3.3, \sigma = 1.0, n = 64$; $\mu_{1-3} = 3.4, \sigma = 1.0, n = 23$), whereas those with more experience rated it lower ($\mu_{4,5} = 2.9, \sigma = 1.3$, $n = 16$). These findings are consistent with the ones in [29], indicating that participants with previous experience of language cafés may have higher expectations on the content of the conversation practice.

When partitioning into experience level and robot settings, the number of respondents gets low in each group, but some possible differences were nevertheless observed for Learning:

Learners without previous experience rated Narrator and Interlocutor highest ($\mu_0 = 3.4; 3.4, \sigma = 0.86; 1.2, n = 18; 15$) and Facilitator lowest ($\mu_0 = 3.2, \sigma = 1.2, n = 16$).

Learners with little previous experience preferred the Interviewer setting ($\mu_{1-3} = 3.7, \sigma = 1.2, n = 6$) and rated Interlocutor lowest ($\mu_{1-3} = 3.2, \sigma = 0.98, n = 6$).

Learners with much experience rated Interviewer highest and Facilitator lowest ($\mu_{4,5} = 3.7; 2.5, \sigma = 2.4; 1.0, n = 3; 4$), but the low number of respondents makes it difficult to draw conclusions (the very large standard deviation for Interviewer is due to one respondent rating all settings except Narrator as 1).

5.10 Gender

Female learners were significantly more positive than male when considering all settings together from a Learning perspective ($\mu_F = 3.4, \sigma = 1.0, n = 62$; $\mu_M = 3.1, \sigma = 1.2, n = 41$, one-way ANOVA, $p < 0.05$). In addition, repeated single factor ANOVA for gender showed significant differences for Friendliness ($p < 0.01$), Personal ($p < 0.05$), Conversing ($p < 0.01$) and Human-like ($p < 0.05$).

There were further gender differences regarding preferences for robot interaction style. Women rated Interlocutor ($n = 15$) highest for Learning, together with Interviewer (both $\mu_F = 3.6, \sigma = 1.1$), and Interlocutor as most Friendly ($\mu_F = 3.7, \sigma = 0.9$). They rated Narrator ($n = 15$) lowest for Learning ($\mu_F = 3.2, \sigma = 1.0$), Friendliness ($\mu_F = 2.8, \sigma = 1.1$) and Personal ($\mu_F = 2.7, \sigma = 1.2$).

Men, on contrary, rated Narrator ($n = 13$) highest for Learning ($\mu_M = 3.5, \sigma = 0.9$), Friendliness ($\mu_M = 3.2, \sigma = 1.0$) and Personal ($\mu_M = 2.7, \sigma = 0.9$), and rated Interlocutor ($n = 10$) the lowest for Learning ($\mu_M = 2.7, \sigma = 1.3$), Friendliness ($\mu_M = 2.3, \sigma = 1.1$) and Personal ($\mu_M = 2.8, \sigma = 0.9$).

As gender hence seems to influence the preference of robot settings, we repeated the analysis of correlation between perceived interaction style and rating (Sect. 5.5), but separately for the two genders, with the results presented in Table 7. We first observe, from the highly significant negative correlation between the amount of Narrating and Learning rating, that male participants who rated Narrator higher did apparently not consider the robot to be Narrating. Indeed, the sessions that the male participants perceived as containing the most robot Narrating were, in descending order, sessions with Facilitator (100% of the session, Learning rating: 1), Narrator (79%, Learning: 4), Interviewer (78%, Learning: 1), Interlocutor (50%, Learning: 4), Narrator (50%, Learning: 3), Interlocutor (45%, Learning: 1), Interviewer (45%, Learning: 3), and male participants hence found that the robot was predominantly Narrating also in other settings than Narrator.

For male participants, we observe that there is a highly significant positive correlation between the proportion that the robot spent on Facilitating and the rating of Learning. However, when analyzing which sessions male participants considered that Facilitating was an important interaction style (11 sessions with proportions ≥ 0.25), 45% were in fact Narrator, 36% Interviewer and 8% each Facilitator and Interlocutor (the corresponding proportions for female participants were 21%, 25%, 29% and 25%, respectively, for 28 sessions) and this positive correlation in fact hence relates more to positive male ratings of Narrator sessions. Further, there was a highly significant positive correlation between the proportion spent on Interview1 and the men's rating of Conversing behavior.

For female participants, there were significant negative correlations between the proportion of Narrating and the rating of how Personal the robot was, its Conversing behavior and it Human-likeness, hence underlining the female participants' more negative impression of Narrating.

Considering gender differences further, we observe differences in correlation between (in each pair the first received higher ratings from men and the second from women):

- Friendliness: Narrating vs. Facilitating;
- Initiative: Interview1/Facilitating vs. Narrating;
- Conversing: Interview1 vs. Participating; and
- Human-likeness: Narrating vs. Interview1/Facilitating.

5.11 Peer Familiarity

Peer familiarity was pooled into the four categories peer unknown (0, $n_u = 28$), peer little known (1–2, $n_l = 32$), peer familiar (3, $n_f = 24$) and peer well-known (4–5, $n_w = 19$).

Table 7 Correlation between the session share of different robot behaviors, as perceived by the subject, and the rating of the session, considering the two genders separately

	Interview1		Narrating		Facilitating		Participating	
	F	M	F	M	F	M	F	M
Learning	0.14	0.18	**−0.24**	**−0.44****	*0.04*	*0.36***	0.08	0.13
Friendly	0.08	0.02	*−0.21*	*0.07*	**0.08**	**−0.27**	0.07	0.02
Personal	0.21	0.21	−0.33*	−0.33*	0.02	0.06	0.10	0.21
Initiative	*−0.09*	*0.22*	**−0.06**	**−0.39***	*−0.01*	*0.22*	0.21	0.16
Conversing	*0.17*	*0.40***	−0.26*	−0.20*	0.05	0.03	**0.04**	**−0.14**
Human-like	**0.23**	**0.09**	*−0.48***	*−0.12*	**0.17**	**0.01**	0.11	0.08

Cell highlight gender differences, with bold indicating higher correlation for female (F) and italic for male (M) participants
Level of significance: $*p < 0.05$; $**p < 0.01$

The ratings were similar over all settings, with the main exceptions being:

For Facilitator, peers who were well-known to each other rated Learning higher than those who were only familiar to each other ($\mu_w = 3.3$ vs. $\mu_f = 2.9$; $\sigma = 1.6; 0.38, n = 3; 7$). The other settings instead have a clear drop in rating for peers who were well-known to each other compared to familiar or little known: Interviewer ($\mu = 4.3 \rightarrow 3.6; \sigma = 0.96; 1.3, n = 4; 7$), Narrator ($\mu = 3.6 \rightarrow 2.8; \sigma = 1.0; 0.60, n = 11; 4$) and Interlocutor ($\mu = 3.5 \rightarrow 3.0; \sigma = 1.2; 1.0, n = 5; 8$).

We interpret these observations, cautiously as the number of responses per category is low, as increased willingness to take more initiative and a wish to interact more with the peer, when this peer is well-known.

5.12 Cultural Origin

Due to the large variety of country of origin, a very coarse grouping was made into the six categories Middle East (Egypt, Iran, Iraq, Kurdistan, Syria, n = 36), Europe (Albania, Croatia, Italy, Poland, Spain, n = 26), Euro-Asia (Afghanistan, Azerbaijan, Kazakhstan, Ukraine, n = 18), Asia (China, Philippines, n = 8), Africa (Congo, Somalia, Eritrea, n = 8), Latin-America (Chile, Cuba, n = 8).

For Learning, the Latin-American subjects were overall the most positive ($\mu_{LA} = 4.1, \sigma = 0.93, n = 8$), while the European were the least positive ($\mu_{Eu} = 2.8, \sigma = 1.1, n = 26$). The African subjects found the robot to be the most Friendly ($\mu_{Af} = 3.8, \sigma = 0.92, n = 8$), and the European the least ($\mu_{Eu} = 2.9, \sigma = 1.3, n = 26$). The Middle-East learners found the robot to be the most Personal ($\mu_{ME} = 3.1, \sigma = 0.94, n = 36$), and the European the least ($\mu_{Eu} = 2.2, \sigma = 1.5, n = 26$). The Asian learners rated the robot's Conversing behavior the highest ($\mu_{As} = 3.5, \sigma = 0.53, n = 8$), and the European the lowest ($\mu_{Eu} = 2.4, \sigma = 1.0, n = 26$).

Only Europe, Middle-East and Euro-Asia were included in the breakdown per robot setting, as the other categories contained too few subjects. The results, summarized in Table 8, show some agreement between groups (e.g., Interviewer being considered the most Personal by both European and Middle-East learners), but also important disagreement (e.g., Narrator being rated highest for Learning by European learners, but lowest by Middle-Eastern), indicating that learner origin (which possibly also includes differences in familiarity with robots and technology enhanced learning) is a factor that may need to be taken into account when setting up the robot moderator.

6 Limitations

A number of limitations of the present study must be acknowledged.

First and foremost, as each learner only interacted with the robot for a total of 40–50 min, few conclusions may be drawn for long-term practice, regarding effectiveness, the learners' preferences for and adaptation to different robot interaction styles or the extent to which the robot can maintain learner interest with sufficiently varied topics for conversation. As the field of robot-assisted language learning is new and both educational robots and RALL methodology are still very much in the development phase, a large majority of previous work has also focused on short-term effects. Such studies are valuable to guide future development, which should then be evaluated with long-term studies.

Secondly, the survey evaluation has two weaknesses. As pointed out above, the subject drop-out, which lead to an imbalanced dataset, and the large variability in participant variables (age, origin, proficiency level) hindered a proper multivariate analysis. Further, survey responses in general provide a quantitative, but rather superficial, evaluation of how the learners' perceived the practice. In one follow-up study, we have therefore interviewed learners post-session to get more qualitative feedback, and in another, we have made quantitative and qualitative analyses of how the robot's

Table 8 Influence of region of origin on preferences on robot setting

	Europe		Middle-East		Euro-Asia	
	Highest	Lowest	Highest	Lowest	Highest	Lowest
Learning	**Narrator**	*Interlocutor*	*Interviewer*	**Narrator**	*Interviewer*	*Interlocutor*
Friendly	**Interlocutor**	**Interviewer**	**Interviewer**	Narrator	Facilitator	**Interlocutor**
Personal	*Interviewer*	**Narrator**	*Interviewer*	Facilitator	**Narrator**	Interlocutor
Initiative	**Interviewer**	Others	*Facilitator*	Interlocutor	*Facilitator*	**Interviewer**
Conversing	**Interlocutor**	*Narrator*	*Interviewer*	**Interlocutor**	*Interviewer*	*Narrator*
Human-like	*Interlocutor*	*Facilitator*	*Interlocutor*	*Facilitator*	Narrator	Interviewer

The highest and lowest rated setting is given for each of the three regions. Agreement between two groups is indicated in italics, disagreement in bold. 'Others' for Initiative-Europe-lowest is due to a tie between the three remaining settings

interaction style influences the interaction with and between the learners.

A number of further limitations (such as the use of a semi-automated wizard-of-Oz setup, that one single interaction style was employed throughout the session, that there was no between-session progression in the conversations between robot and learners, since the robot had no memory) are discussed in Sect. 8 related to future work to improve the robot–learner interaction.

7 Discussion

The results presented above can be discussed with respect to several different perspectives.

The fact that the participants did not judge the content of the sessions to be very different for the different robot strategies has already been discussed in Sect. 5.4, where it was also shown that there was nevertheless a difference between settings in accumulated preference scores for the completion group, and that the subjects' perception of the distribution between different strategies within the session influenced their ratings. We therefore conclude that the different interaction strategies may, in general, be more or less appropriate for L2 conversational practice. However, we also identified clear differences between different learner categories (L2 level, gender, experience of language cafés, familiarity with the peer, cultural origin), which signifies that the interaction strategy needs to be adapted to the learners, as discussed in Sect. 8.1.

In relation to previous work in RALL, we observed, just as e.g., [38] that learners reported that practicing with a robot was less intimidating than with a human teacher, that they were interested in exchanging personal information with the robot (c.f. further Sect. 8.2) just as the learners in [17], and that the setting with two learners resulted in linguistic- or topic-related peer collaboration, similar to the studies with a robot peer [13,14,20]. The present study differs from earlier work in that the collaborative setting is with two adult human learners. As discussed in Sect. 2.1, the collaborative setting is important to incorporate peer support, and we observed how learners in fact did rely on each other to tackle problems understanding the TTS or finding words for their own utterances. Targeting adult learners has implications in that more realistic robot appearance and behaviour are required than for younger learners. We find that Furhat is well-suited for realistic conversational practice with L2 learners, but that the use of non-verbal displays should be increased or improved, e.g., regarding eye contact, gaze, blinking behaviour and visual emotional display.

Furthermore, the set-up with a robot and two learners could satisfy several learning styles for foreign language learning [39], since it incorporates different aspects from the perceptual, social and cognitive dimensions.

For the perceptual dimension, the primary focus is the **linguistic** and **auditory**, with practice of verbal speaking and listening skills, but also include **spatial** and **kinesthetic** referencing, as the situated interaction encourages learners to use body language, such as physical cues for turn-taking, addressing and referencing, and the robot will (to some extent, since it is a head-only setup) use similar spatial signals. Physical enacting has, e.g., been shown to be beneficial in RALL for learning of verbs [13].

For the social dimension, the setting with two learners allows for **interpersonal**, collaborative language learning, in which the peers support each other [2], in addition to creating a social relationship between the three conversation partners, which has been shown to be effective to maintain learner interest for RALL [17].

The cognitive dimension of learning is more complex, but the realistic social conversations on familiar topics should suit **extrovert concrete sequential** learners and the language café type focus on communication rather than linguistic form **impulsive holistic** learners.

As learners differ in preference along the above dimensions, language café practice may in general be more or less suitable for individual learners, but differences in learning style may also influence preferences for robot interaction

strategies. This was not explored in the present study and such an investigation would be relevant in future work.

Another aspect concerns the benefit of using a robot, as opposed to a voice-only or a screen-based interface. The motivational effect of the robot was described in the Introduction, but more importantly, the embodiment influences the interaction between both the learners and the learning software and the interaction between the learners. In two follow-up studies, we have made quantitative and qualitative analyses of how the robot's strategy influenced the interaction and have interviewed participants on how they perceived the interaction and the robot. The studies showed that (1) the robot's anthropomorphic appearance and behaviour were positively received, (2) it lead to personification of the robot as a social counterpart in the conversation (c.f. further Sect. 8.2), (3) head-turning towards one learner was important for turn-taking, and (4) respondents were in general positive regarding the value of the language practice with Furhat and would like to use it again. Even if we have not yet compared the conversational practice with a robot to those with a smart speaker or a screen-based agent, we hence nevertheless see clear benefits of using robots for the practice.

8 Future Work

We have identified a number of areas for improvement, which we list in Sects. 8.1–8.3 as recommendations for similar conversational practice sessions with a robot, as well as requirements for our own future work.

8.1 Adaptation to Learners

Identify the Target Learner Group We have in this, and our previous study [29], observed that the robot is able to hold a social conversation that is found to be meaningful from a learning perspective by the learners, provided that they themselves have a basic level (B1 or possibly A2) allowing them to engage in basic spoken conversations, and that their level is not too advanced (above B2 or C1, depending on the setting), making them demand more flexible and complex conversations.

Similarly, we observed in both studies that learners with more experience with standard language cafés have higher expectations on the conversations and therefore rated the sessions with the robot lower.

Hence, at the present time, the target group for robot language cafés, is primarily B1 learners with less possibilities of participating in regular language cafés. However, there is currently an international effort taking place in conversational artificial intelligence, with the goal to create social bots able to e.g., *"converse coherently and engagingly with humans on a range of current events and popular topics such as entertainment, sports, politics, technology, and fashion"*.[8] It is hence probable that robots in the near future will be able to converse socially at a far more advanced level, but this would make session-wise adaptation to learners necessary.

Choose Robot Interaction Style Based on the Learners' Relation We did find some support for the assumption that participants who know each other well want to interact more with each other, rather than letting the robot control the conversation. Participants who were less familiar to each other did on the other hand to larger extent want the robot to keep the initiative in the dialogue. Determining, e.g., as a part of initial social introductory questions from the robot, if the participants already know each other, and adapt the interaction style based on this, is hence recommended.

Choose Robot Interaction Based on Learners' L2 Level In addition to the general adaptation of moderator interaction to the participants, in terms of e.g., speaking rate and complexity of utterances, we observed in this study that learners at different levels may want the robot to interact differently.

In general, our findings are in line with what could be expected, from experience in the L2 classroom and the survey answers by human moderators (c.f. Sect. 3.1), i.e., that more advanced and experienced learners will interact more with each other and take more initiative to propose topics and ask the moderator questions. Adaptation to learner level could be made e.g., based on a pre-session self-rating or by monitoring the session (number of learner requests for clarification or repetition, length, speaking rate and ASR confidence scores of learner utterances).

Such increased engagement requires that the robot's interaction capabilities are improved.

8.2 Improvement of the Robot'S Interaction Capabilities

Build and Access Database with Robot Responses to User Questions In our previous study [29], we observed that learners would frequently ask the robot questions, to socialize (returning the same type of questions that the robot asked them), to learn more about the robot (asking for information about its technical properties and functioning) or to test the system (trying to determine its conversational capabilities). In that study, such questions were handled by allowing the wizard to use ASR to generate appropriate answers. This, however, lead to very long response times for the robot.

For this study, we instead created a substantial number of utterances covering the "personal background" of the

[8] Alexa challenge, http://developer.amazon.com/alexaprize.

robot (hobbies; food, travel, film, book and sports preferences; family etc), technological information about the robot (components, creators, history) and a "personality" (jokes, non-controversial political beliefs, social responses). However, due to the scope of the study (i.e., to investigate each "pure" interaction style) these answers were only accessible in the Narrator and Interlocutor settings. Moreover, due to the current implementation, in which the transitions from each state are pre-defined, the robot could only answer such questions that had been anticipated for the present state. If a participant asked a question to which there in fact existed a pre-generated answer somewhere else in the dialogue tree, the wizard could not select this utterance. We found that this negatively affected the ratings by the more advanced participants, as they to larger extent asked the robot questions and would often not get a relevant answer. As an illustration of this, one participant noted in the free text field after the Interviewer session *"Only the robot asked questions"* and another after the Interlocutor session *"This time the robot was more interesting. He responded to our questions"*.

For the robot to be credible beyond this type of study of "pure" interaction styles, it is necessary that it should have an utterance repertoire enabling it to answer at least a basic set of social questions (of the same type it is asking the participants) and provide information about itself, *regardless* of the present dialogue state.

Transparency in What Topics That the Robot is Able to Discuss Some of the lower ratings for Facilitator are probably linked to the fact that the robot encouraged the learners to suggest a topic, but it was then not able to provide any own input on that particular topic, if it was outside the domain of predefined topics. The intention was that the two participants should start talking about their suggested topic with each other, but they instead often expected the robot to lead the conversation. In order to avoid such problems, the participants must know what topics the robot can discuss. For settings like Interviewer, Interlocutor and Narrator, this is fairly easy, as the robot is mostly maintaining the initiative in the dialogue, steering it to topics that it can handle, but for Facilitator-type dialogues, it is appropriate that the participants are told, explicitly or implicitly, what topics the robot is interested in. Transparency is a common HRI requirement to ensure that the users interacting with a robot are able to build a mental model of the robot's capabilities. In the language learning setting, transparency is of even higher importance, since the learners' own interaction capabilities are impeded, due to limited language skills.

Robot Memory is Required for Long-Term Interaction If a learner has already interacted with the robot, she will firstly expect a socially competent robot to remember her previous answers (note the remark in Section 5.4 about already having told the robot) and secondly not appreciate hearing the same personal information about the robot again. Ideally, the conversation should also be personalized for the learner, so that the selection of topics is based on previous learner input. The importance of personalization for long-term interaction has been demonstrated previously [17].

Consequently, a robot memory needs to be generated to keep track of what topics have already been discussed with a returning learner, and what information the learner provided in her previous responses. Topics that have already been discussed in a previous session could then be pruned when loading the dialogue tree for a new session. Our preliminary work, using tf-idf and word2vec vectorization of the utterances, shown promising results of avoiding repetition by determining similarity with questions that have already been asked. Personalizing the dialogue and robot utterances to previous learner input is more complex, as it requires that a mapping between learner information and topics and particular utterances needs to be learned.

Mix Interaction Styles We should emphasize that the aim of this study was to be able to study the different interaction styles separately, not that any of them would be *the* interaction style to choose. As illustrated clearly by Figs. 1 and 4, most human moderators switch between different interaction styles within the same session, and a robot moderator should naturally also do this.

Refine Interaction Styles Just as L2 learners are improving their conversational skills in the language café setting, so should the robot, automatically through machine learning or manually by updating utterances and possible transitions, based on observations of how successful the respective dialogues were. The main required refinement of the current settings are:

Narrator was rated highly when it included a higher amount of interaction with the learners (quiz, chit-chat that included asking for the participants' views) and shorter robot turn lengths, whereas sessions with verbose robot utterances were rated poorly. We will therefore simplify and shorten the robot turns and use Narrator for shorter sub-dialogues within other settings.

Facilitator will be used as a complement to the Interlocutor setting, rather than on its own. It is natural to start the dialogue with robot initiative, but then switch to a Facilitator strategy, if it is determined that the participants are more interested in interacting with each other.

Interviewer and *Interlocutor* will be pursued as settings, separately or in combination. Interviewer would benefit from being more personal and allowing for more multi-directional interaction (one subject remarked *"He asked most of the questions, it was sort of like an interview."*), which could be achieved by introducing Interlocutor properties. Interlocutor could on the other hand benefit from "keeping it simple" and also include simpler interaction similar to Interviewer. The utterances and allowed transitions in the current Interlocutor setting were formulated so as to increase the differences

compared to Interviewer, thus sometimes making the dialogue overly complicated, at least for lower level learners.

8.3 Technological Development

We end by considering some technological aspects that we will address in our future work.

Adaptive Speech Synthesis Speed Eight subjects wrote in the survey that the robot talked too fast, and several in addition asked the robot during the session if it could talk more slowly. Being able to automatically adjust the the speech synthesis speed to the complexity of the utterance, and to repeat it more slowly, if the participants ask for it or fail to understand, is hence essential. In preliminary work, we have performed complexity analysis of the robot utterances—regarding word and trigram frequencies in the sentences—as a means to determine which utterances, and which parts thereof, may need a reduced speaking rate.

Autonomous State Transitions We will use deep learning on the data of the dialogues collected in this study, reinforced by the survey ratings and our own hand-labeling, to train the robot to autonomously select the best response, given the current and previous states and the user utterance. Our mark-up will focus on if the selected robot utterance was apt (positive reinforcement), neutral or inapt (negative reinforcement) for the progress of the dialogue, while the survey ratings will be employed to identify successful dialogues to use as models. Using this off-line input, we will train a neural network, based on sequence-to-sequence modeling [40], or a hybrid code network [41] employing a recurrent neural network combined with domain knowledge, to manage the dialogue flow and take into account the history of the dialogue. Such networks would further allow connecting a previous dialogue with the same learner to the entry node of the current dialogue, hence enabling the robot to remember previous sessions with the learner.

Automatic Speech Recognition for L2 Learners in Conversations We have also begun to explore the recorded audio data to attempt to increase the robustness of the standard ASR, by using mapping between off-line ASR on the recorded learner utterances and manual labeling of the learner's intent for the same utterances. This procedure will allow for machine learning of a post-processing stage, in which the output from the ASR, when run in real-time during future sessions, is corrected to align with the learner's intent, using information of the current dialogue state (which gives probabilities for plausible learner utterances). Note that this makes the complex task of speech recognition for beginner-level L2 learners more tractable, since keywords and concepts need to be recognized, rather than the full utterance verbatim.

Audiovisual Detection of and Adaptation to Learner Engagement We have further initiated preliminary work on using the video recordings of the learner's face expressions and body posture, and the audio recordings of each learner's voice, to train a module that is tracking how the learners perceive the current state of the dialogue.

We have investigated if the post-session rating of the learning and the dialogue interaction can be predicted from automatic audiovisual analysis of the recorded session, using standard technology for facial feature tracking (OpenFace [42]) or detection of frustration or disengagement in vocal features (OpenVokaturi SDK[9]). The correlation between the audiovisual data and our collected survey responses was weak. However, using manual annotation of the audiovisual data, machine learning was successful at estimating engagement of the active speaker and the listener. In future sessions, we will use the audiovisual tracking to determine when the robot's interaction needs to be changed within the session, in order to improve it from the learner's perspective.

9 Conclusions

We have in this study explored the large repertoire of interaction strategies of human language café moderators and how these strategies are influenced by different learner variables. Based on this exploration, we implemented four stereotypic interaction styles in a social conversational robot and performed an on-site user study with Swedish for immigrants students. The responses from the participants in the user study and our own observations resulted in a combined list of recommendations for conversational L2 practice with robots and specific improvements that we will address in our future work. As we are quite confident that educational conversational robots will proliferate in the years to come, we believe that this study can be of interest to the spoken human-robot interaction community in general and to researchers working on robot-assisted language learning in particular. We will continue to develop our robot's conversational interaction skills and will iteratively test it with both learners similar to the ones in this study and other target groups (e.g., children or for professional-domain-specific interactions).

Acknowledgements Open access funding provided by Royal Institute of Technology. The authors would like to thank coordinators and moderators of language cafés around Sweden for providing valuable input to the survey on human moderator strategies; and directors and students of Hermods' Swedish for Immigrant classes for respectively welcoming the user study on their premises and participating in the study. Duy M. Nguyen, CoRIS Institute, Oregon State University, proposed the correlation analysis in Sects. 5.5 and 5.10.

Funding This study was funded by Swedish Research Council (Grant 2016-03698 "Collaborative Robot Assisted Language Learning (CORALL)").

[9] https://developers.vokaturi.com/.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Spada N (2007) Communicative language teaching—current status and future prospects. Springer, Berlin, pp 271–288
2. Nunan D (1992) Collaborative language learning and teaching. Cambridge University Press, Cambridge
3. Leite I, Martinho C, Paiva A (2013) Social robots for long-term interaction: a survey. Int J Social Robot 5(2):291–308
4. Alemi M, Meghdari A, Ghazisaedy M (2015) The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition. Int J Social Robot 7(4):523–535
5. Neri A, Cucchiarini C, Strik H (2002) Feedback in computer assisted pronunciation training: technology push or demand pull? In: Proceedings of CALL conference "CALL professionals and the future of CALL research, pp 179–188
6. Han J (2012) Emerging technologies, robot assisted language learning. Lang Learn Technol 16(3):1–9
7. Aidinlou NA, Alemi M, Farjami F, Makhdoumi M (2014) Applications of robot assisted language learning (RALL) in language learning and teaching. Int J Lang Linguist 2(3–1):12–20
8. Mubin O, Shahid S, Bartneck C (2013) Robot assisted language learning through games: a comparison of two case studies. Aust J Intell Inf Process Syst 13(3):9–14
9. 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
10. Han J, Jo M, Park S, Kim S (2005) The educational use of home robots for children. In: Proceedings of the 14th IEEE international workshop on robot and human interactive communication, pp 378–383
11. 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 interaction (HRI '16), pp 231–238
12. Khalifa A, Kato T, Yamamoto S (2016) Joining-in-type humanoid robot assisted language learning system. In: Proceedings of LREC, pp 245–249
13. Tanaka F, Matsuzoe S (2012) Children teach a care-receiving robot to promote their learning: field experiments in a classroom for vocabulary learning. J Hum Robot Interact 1(1):78–95
14. Mazzoni E, Benvenuti M (2015) A robot-partner for preschool children learning English using socio-cognitive conflict. Educ Technol Soc 18(4):474–485
15. Balkibekov K, Meiirbekov S, Tazhigaliyeva N, Sandygulova A (2016) Should robots win or lose? Robot's losing playing strategy positively affects child learning. In: Proceedings of IEEE international symposium of robot and human interactive communication, pp 706–711
16. Kanda T, Hirano T, Eaton D, Ishiguro H (2004) Interactive robots as social partners and peer tutors for children: a field trial. Hum Comput Interact 19(1):61–84
17. 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 Robot 23(5):962–971
18. Admoni H, Scassellati B (2014) Roles of robots in socially assistive applications. In: IROS 2014 workshop on rehabilitation and assistive robotics
19. 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
20. 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: Proceedings of ISCA workshop on speech and language technology in education, pp 13–17
21. Wedenborn A, Wik P, Engwall O, Beskow J (2016) The effect of a physical robot on vocabulary learning. In: Proceedings of the international workshop on spoken dialogue systems
22. Morton H, Jack MA (2005) Scenario-based spoken interaction with virtual agents. Comput Assist Lang Learn 18(3):171–191
23. Wik P, Hjalmarsson A (2009) Embodied conversational agents in computer assisted language learning. Speech Commun 51(10):1024–1037
24. Johnson L, Valente A (2008) Tactical language and culture training systems: using artificial intelligence to teach foreign languages and cultures. In: Proceedings of AAAI, pp 1632–1639
25. Hautopp H, Hanghøj T (2014) Game based language learning for bilingual adults. In: ECGBL2014-8th European conference on games based learning: ECGBL2014. Academic Conferences and Publishing International 191
26. Åhlund A, Aronsson K (2015) Stylizations and alignments in a L2 classroom: multiparty work in forming a community of practice. Lang Commun 43:11–26
27. Åhlund A, Aronsson K (2015) Corrections as multiparty accomplishments in L2 classroom conversations. Linguist Educ 30:66–80
28. Cekaite A, Aronsson K (2005) Language play, a collaborative resource in children's L2 learning. Appl Linguist 26(2):169–191
29. Lopes J, Engwall O, Skantze G (2017) A first visit to the robot language café. In: 7th ISCA workshop on speech and language technology in education (SLaTE), pp 7–12
30. Kipp M (2012) Multimedia annotation, querying and analysis in ANVIL. Wiley, Hoboken, pp 351–368
31. Jonell P, Bystedt M, Fallgren P, Dimosthenis K, Lopes J, Malisz Z, Mascarenhas S, Oertel C, Raveh E, Shore T (2018) Farmi: a framework for recording multi-modal interactions. In: Proceedings of LREC, pp 3969–3974
32. Skantze G, Al Moubayed S (2012) IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In: Proceedings of ICMI
33. Al Moubayed S, Beskow J, Skantze G, Granstrom B (2012) Furhat: a back-projected human-like robot head for multiparty human–machine interaction. In: Esposito A, Esposito AM, Vinciarelli A, Hoffmann R, Müller VC (eds) Cognitive behavioural systems. Springer, Berlin, pp 114–130
34. Skantze G, Johansson M, Beskow J (2015) Exploring turn-taking cues in multi-party human–robot discussions about objects. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, pp 67–74

35. Johansson M, Beskow J (2015) A collaborative human-robot game as a test-bed for modelling multi-party, situated interaction. In: In Lecture notes in artificial intelligence, Proceedings of intelligent virtual agents, pp 348–351
36. Jonell P, Mendelson J et al (2017) Machine learning and social robotics for detecting early signs of dementia. arXiv:1709.01613v1
37. Kanov M (2017) Sorry, what was your name again? How to use a social robot to simulate alzheimer's disease and exploring the effects on its interlocutors. KTH Royal Institute of Technology, Stockholm
38. 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, vol 9388
39. Oxford RL (2003) Language learning styles and strategies: an overview. In: Proceedings of generative approaches to language acquisition, pp 1–25
40. Vinyals O, Le QV (2015) A neural conversational model. arXiv:1506.05869
41. Williams JD, Asadi K, Zweig G (2017) Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. arXiv:1702.03274
42. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings of conference on computer vision and pattern recognition, pp 815–823

Olov Engwall is professor in Speech Communication at KTH Royal Institute of Technology. He received his Ph.D. in 2002 with a thesis on multimodal articulatory speech production modeling and has since focused e.g., on computer-animated tutors for pronunciation training. He leads the Swedish research project on Collaborative Robot-Assisted Language Learning.

José Lopes is research associate at the Interaction lab at Heriot-Watt University. He received his Ph.D. 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.

Anna Åhlund is senior lecturer at the Section for Early Childhood Education at Stockholm University. Her Ph.D. thesis from 2015 focused on collaboration in the second language classroom and her continued research interest is in second language learning of Swedish.