7th ISCA Workshop on Speech and Language Technology in Education
25-26 August 2017, Stockholm, Sweden

A First Visit to the Robot Language Café

José Lopes, Olov Engwall, Gabriel Skantze

KTH Royal Institute of Technology, Sweden
jdlopes@kth.se, engwall@kth.se, skantze@kth.se

Abstract

We present an exploratory study on using a social robot in a conversational setting to practice a second language. The practice is carried out within a so called language café, with two second language learners and one native moderator; a human or a robot; engaging in social small talk. We compare the interactions with the human and robot moderators and perform a qualitative analysis of the potentials of a social robot as a conversational partner for language learning. Interactions with the robot are carried out in a wizard-of-Oz setting, in which the native moderator who leads the corresponding human moderator session controls the robot. The observations of the video recorded sessions and the subject questionnaires suggest that the appropriate learner level for the practice is elementary (A1 to A2[1]), for whom the structured, slightly repetitive interaction pattern was perceived as beneficial. We identify both some key features that are appreciated by the learners and technological parts that need further development.

Index Terms: conversational L2 practice, human-robot interaction, computational paralinguistics

1. Introduction

Developing conversational skills in a new language can certainly be a challenging task for an adult second language (L2) learner. Other skills, such as vocabulary, grammar and written and oral comprehension, can be readily practiced, in the L2 classroom, or with self-study exercises and by passive exposure. In contrast, it is difficult to allocate an adequate amount of time for individual practice of conversational skills in an L2 classroom with many students, and learners' progress is therefore very much dependent on their practice outside the classroom. For adult immigrants learning Swedish, this has been shown to be difficult, for two reasons:

Firstly, Sweden has welcomed a number of refugees over the past decades that is very large in proportion to the number of inhabitants (e.g., in 2007, the town of Södertälje, with 80.000 inhabitants, welcomed more Iraqi immigrants than the US and Canada together), and many of these immigrants live in communities where they mainly interact with compatriots. Secondly, almost the entire Swedish population is proficient in English as an L2, with 86% self-reporting that they can hold a conversation in English [1], and it is hence extensively used as lingua franca for social conversations.

However, even if English is widely accepted for social interactions, integration in society in general and on the job market in particular is very dependent on the mastery of Swedish. To complement the traditional classroom teaching, many societal agencies (NGAs, churches, universities) therefore arrange Language Cafés to help learners practice conversation.

A Language Café is an open gathering to which native and non-native speakers convene spontaneously to engage in open-topic social conversations for an hour or so, with the explicit aim of maintaining the conversation in the target language, even in case of linguistic problems. The native speakers are often given the extra responsibility of initiating the conversations and assisting the L2 learners with linguistic problems, but they nevertheless have the role as equal conversational partner, rather than as a teacher. The language cafés therefore constitute a relaxed setting, in which errors, hesitation and slow interaction are allowed to a far greater extent than in everyday interaction. However, there is often a shortage of native speakers for the language cafés, in general, and in immigrant communities and at asylum residences in particular. Technology-enhanced learning solutions are therefore an attractive alternative to be able to provide more L2 learners with practicing opportunities.

Over the last decades, computer-assisted language learning software has been developed to practice communication skills. Systems such as SPELL [2], DEAL [3], The Tactical Language and Culture Training System [4] and Dansksimulatoren [5] allow learners to practice spoken interaction in 3D virtual reality environments, using respectively, task scenarios such as ordering meals at the restaurant, bargaining at a flea market, interacting with Iraqi citizens as a US army soldier, and exploring a Danish community as an immigrant. While these systems may provide valuable training time, the on-screen virtual setting fails to practice real world skills such as multimodal turn-taking and communication signals.

Within a newly initiated research project on collaborative robot-assisted language learning (RALL), we have therefore started to explore how a social anthropomorphic robot could be used as native speaker in a language café setting with two L2 learners of Swedish. Language café interactions can vary greatly depending on the learners' level of Swedish, but the concept of mostly meeting with new persons of unknown background and linguistic level, entails that the topics of the conversations are often rather limited and repetitive. They focus on personal aspects (occupation, family, hobbies, food and travel preferences) or comparisons between different countries, cultures and languages. While this may perhaps be a downside for the returning visitor to a language café, it makes the use of a robot as a native conversation leader more achievable: since the expected number of topics is limited, the robot's dialogue system can control the interaction more and have better chances of predicting answers from the conversational partners. This is important as the input from the learners can be expected to contain a substantial amount of errors of vocabulary, grammar and pronunciation, which makes the task for the automatic speech recognition (ASR) challenging if the topics are wider. The language café setting comes with an additional benefit for RALL, since the two peers may support each other, both to correct erroneous utterances that could not be understood by the robot, and to monitor if a non- or mis-understanding by the robot of something that the peer says is due to actual linguistic problems or in fact to technological shortcomings of the system. This is

[1] European Common Language Reference Framework

10.21437/SLaTE.2017-2

also of importance, as erroneous feedback from a CALL system, either explicit or in terms of misrecognition of a correctly produced utterance, may be detrimental for learning [6].

Developing a robot that can participate in a multi-party, multimodal, open-domain conversation is, to our knowledge, a new challenge in RALL. In this paper we present a first step in that direction recording recorded a corpus of language café interactions, both with a human native speaker and a (wizard-of-Oz controlled) Furhat robot head leading the conversation.

We are aiming to address four main questions in this exploratory study: 1) to compare the interactions in the two conditions in general, and the effect of the order in which they interacted with the human or the robot moderator. 2) to assess if there is an ideal level of the learners for this type of exercise. 3) to find out technological improvements required for a successful autonomous interaction of the robot in this setting. 4) on the meta level, to investigate to what extent the experience of the wizard (confederate versus naïve) influenced the interaction.

2. Related Work

Educational robot tutors for language learning is still a very recent research field, which has emerged during the last decade (c.f. [7] and [8] for surveys). Previous work in robot-assisted language learning has almost exclusively targeted school and pre-school children and have mainly focused on the motivational aspects of language learning with robots. Examples show that children's English speaking skills as well as confidence and motivation increased significantly through interaction with robot tutors [9], and that robots can be used in game settings with children to teach a foreign language [10].

The robots have either employed a video screen to allow for telepresence display of a human teacher, or have been toy-like robots, such as Philips iCat [10] or the yellow snow-man Keepon [11]. The recently launched European project L2tor uses humanoid Nao robots that interact with pre-school children one-to-one to teach them English, Dutch or German [12]. Nao robots can use human-like body gestures, but can display neither extra-linguistic human facial expressions conveying emotion and turn-taking signals (e.g., smile, eye and eyebrow movements) nor lip movements to give linguistic information, which is fundamental in spoken L2 learning. The anthropomorphic robot Furhat (c.f. Section 3) used in the current study has clear benefits in that it can mimic a human interlocutor.

3. Experimental Setup

The experimental setup for this data collection is shown in Figure 1. The full scene was recorded with one digital video camera and in addition each participant was recorded individually, using a head-mounted microphone for the audio and a dedicated GoPro camera to record facial expressions. The individual audio and video recordings will be used for future training of detection of linguistic errors, hesitation and motivational state. The robot's actions were also logged for further analysis.

Three participants were invited to each experiment session; one fluent Swedish speaker and two L2 learners of Swedish; and were told that they were going to have a language café type of interaction. For those who were not familiar with language cafés, they were instructed to engage in a social conversation on any topic of their choice. The session consisted of two 15-minute interactions, either starting with the L1 speaker as moderator and then switching to Furhat as moderator, or vice versa. The moderator was instructed to lead the conversation and to help the L2 learners when they so require, but the L2 speakers were also free to initiate topics of their interest and to assist each other with linguistic difficulties.

3.1. The Furhat social robot

The robotic head used, Furhat [13], can display a wide variety of human face gestures, as it uses a computer-animated face projected on a 3D mask (c.f. Figure 1). As the neck is fitted with a motor-servo, Furhat can also nod and turn the head, which has been shown to be important in three-party interactions [14]. The robot is controlled by the IrisTk framework [15], which interfaces to components for text-to-speech synthesis, facial animation, automatic speech recognition and interaction event tracking.

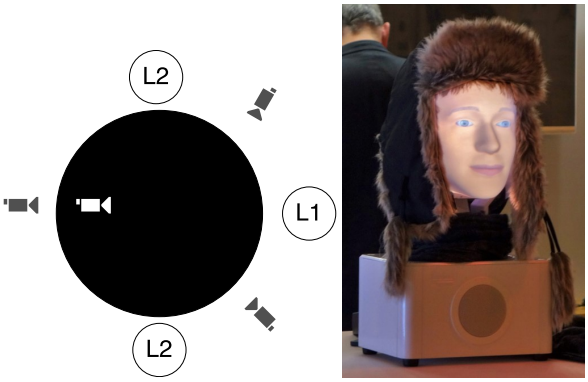

Figure 1: *Experimental Setup and the Furhat robot.*

3.2. Wizard-of-Oz control

The Furhat robot was controlled by the same native speaker who acted as moderator in the setting with a human moderator. Whenever the moderator was not a confederate, (s)he was instructed separately before the session and was given a (very) short introduction to the wizard interface. In this interaction, the wizard was sitting in an adjacent, sound isolated control room, and could observe the language café scene with Furhat and the two learners through a VoIP connection with sound and video. The wizard could control the robot's interaction in the following manner:

Spoken output: Through 1) *Buttons* with pre-generated greetings, questions and responses that are commonly occurring in language cafés sessions; or 2) *ASR input* allowing the wizard to provide other utterances. These were first recognised using Google Cloud ASR and presented on the wizard's screen. When the recognized text corresponded to the wizard's intention, (s)he pressed a button and the utterance was synthesized and spoken by Furhat. While waiting for the utterance recognised, Furhat would signal the intention to speak by uttering a filled pause when the ASR was activated by the wizard.

Non-verbal communication (controlled by buttons): 1) *Head turns* to adress a particular person to indicate whom a question was directed to. This active action by the wizard overrules the automatic direction-sensitive turning of the robot's head towards the current speaker. 2) *Non-verbal feedback*, e.g., "Mhm", to backchannel when someone else is speaking

4. Data Collection

The subjects for the experiment were recruited from the KTH language cafés, on-line recruitment platforms and confederates, and were told that they were going to be part of an experiment that aimed to collect data to improve a robot that can participate in language cafés. For the L2 learners we recommended a level above A2, but everyone who was interested in practicing Swedish was welcome. Each L2 participant was allowed to take part in the experiment once, whereas the L1 speaker could act as moderator several times (and three of them did).

Once both conversations were concluded the participants were asked to fill in a short questionnaire; one for the L2 speakers and one for the L1 moderator. The questions to the L2 speakers were (unless otherwise stated, answers were on the scale 1-5): *1) Have you ever been to a language café before?* [Yes/No] *2) Self-rated proficiency in Swedish.* [Beginner, Elementary, Intermediate, Proficient] *3,5) I feel that the [human/robot] moderator understood what I was saying. 4,6) I understood what the [human/robot] moderator was saying. 5,7) I feel that the [human/robot] moderator helped me in the interaction. 8) Compared to the interaction with the human moderator, that with the robot was [Scale: Much worse (1) to Much better (10)]. 9) What features need to be improved in the robot for it to be useful in a language café setting?* [Speech understanding, Speech synthesis, Response time, Interaction style, Choice of topics, Other] *10) What features did you like in the interaction with the robot moderator? 11) Additional comments about the practice.* The questions to the L1 moderator focused on the perceived differences between being on-site and mediated through Furhat, and on suggestions for improvement. These questions were of a more qualitative nature and are hence analyzed together with the general initial observations in Section 5.2.

We performed 11 experiment sessions with a total of 22 L2 Swedish learners (7 female and 15 male; average age 29.1 years) and 6 different moderators (3 female and 3 male; average age 36.2 years). For their participation, subjects received a cinema ticket as compensation. Of the 11 experiment sessions, 6 started with a human moderator (condition 1) and 5 started with Furhat (condition 2); 8 dialogues had a male moderator and 3 a female moderator. One of the moderators was a non-native, but fluent speaker of Swedish. She and one other recruited moderator had previous experiences of moderating language café sessions. The 22 participants had 14 different mother tongues; some of them were bilingual and all were proficient speakers of English (which influenced strategies employed to handling linguistic uncertainty, as discussed below).

5. Analysis

The number of varying parameters (L2 level of subjects, session order, wizard experience, familiarity with the language cafés, subject and moderator gender) is too large to allow for a rigorous statistical analysis of the impact of each parameter separately. However, some observations may nevertheless be made, to guide the future development, based on the user questionnaires and/or the video recordings.

5.1. Questionnaires

Table 1 summarizes the comparisons of the questionnaire results. Unsurprisingly, in general, the paired t-test showed there are significant differences in the way the participants felt understood by the human moderator compared to the robot moderator and how they got help from him/her. However, there were no significant differences in the way they could understand the robot and human moderators. When comparing the interaction with the robot moderator to that with the human dito, the average score was 3.9 on a scale from 1 ("Much worse") to 10 ("Much better"), indicating that the interaction with the robot was perceived as inferior to that with the human, but not severely so.

We will divide our analysis in three dimensions that could possibly affect the interaction and experience of it: 1) the self-reported language proficiency, to see if a user proficiency level is more suitable than others, 2) if the users have previous experience of language cafés, to see if this influenced the expectations and 3) if the wizard is a confederate or a recruited subject, to see if previous experience of wizarding tasks affects the overall outcome of interaction.

5.1.1. Self-reported proficiency

The levels that could be reported in the questionnaire varied from Beginner to Proficient, and 4 participants reported that they were beginners, 6 that their knowledge of Swedish was elementary, 11 intermediate and 1 proficient (who was discarded in this analysis). Table 1 shows how the questionnaire answers related to the proficiency level.

The scores related to proficiency level can be analyzed with paired t-tests and summarized as: while intermediate learners preferred the interaction with the human moderator (comparison score 3.6, significant differences in the feeling of being understood by the moderator, understanding the moderator and being helped by the moderator), elementary level learners (comparison score 4.2, non-significant differences in understanding the moderator and being helped by him/her) and in particular beginners (comparison score 5.0, no significant differences in scores for being understood, understanding and being helped by the moderator[2]) were relatively more positive towards the robot moderator.

Comparing between different proficient levels, the statistical test (one-way ANOVA) showed no significant difference between levels with respect to feeling understood by the moderator (robot or a human). Differences were significant when it comes to understanding the robot moderator ($F(3) = 0.733, p = 0.546$) and understanding the human moderator ($F(3) = 9.465, p < 0.001$), with more proficient participants understanding the moderators better. The post-hoc Tukey revealed an effect between Elementary and Beginner ($p < 0.05$) and Intermediate and Beginner ($p < 0.001$). We found significant differences in the way subjects felt helped by the human moderator ($F(3) = 3.846, p < 0.05$), with an effect between Beginner and Intermediate (post-hoc Tukey test, $p < 0.05$). Regarding the feeling of being helped by the robot moderator, there was no significant level difference between levels ($F(3) = 2.624, p = 0.08$). Finally, no significant differences between the interaction with robot or human were found between different levels ($F(3) = 1.412, p = 0.272$).

5.1.2. Previous experience in language cafés

Of the 22 L2 Swedish speakers, 15 reported that they had already been to a language café, whereas the remaining 7 answered it was the first time for them.

The results can be summarized as: Experienced language

[2]Note however that the fact that the differences were not significant for the beginners might also be due to the fact that there were fewer (four) participants at this level.

Table 1: *Average scores for feeling understood by the moderator (scale: 1-5), understanding the moderator (scale: 1-5), feeling helped by the moderator (scale: 1-5) and comparing the interaction (scale: 1-10) with a human (Hum) and robot (Rob) moderator w.r.t. proficiency level, language café experience and wizard experience. Statistical differences are reported as ns=non significant, * for $p < 0.05$, ** for $p < 0.01$ and *** for $p < 0.001$, using paired t-tests*

	General			Proficiency level									Language café expericence						Wizard experience					
				Beginner			Elementary			Intermediate			Experienced			New			Confederate			Recruited		
	Hum	Rob	Δ	Hum	Rob	Δ	Hum	Rob	Δ	Hum	Rob	Δ	Hum	Rob	Δ	Hum	Rob	Δ	Hum	Rob	Δ	Hum	Rob	Δ
Understood by	**4.6**	**3.3**	***	4.0	2.75	ns	**4.7**	**4**	*	**4.7**	**3.3**	***	**4.5**	**2.8**	**	**4.7**	**3.3**	*	**4.7**	**3.3**	*	**4.5**	**3.3**	**
Understanding	4.2	4.1	ns	3.0	3.5	ns	4.2	4.2	ns	4.6	4.2	*	**4.6**	**4.1**	*	4.0	4.2	ns	**4.6**	**4.1**	*	4.0	4.2	ns
Helped by	**4.2**	**2.7**	***	3	2.5	ns	4.0	3.3	ns	**4.7**	**2.5**	***	**4.3**	**2.5**	***	**3.9**	**2.9**	ns	**4.6**	**2.9**	***	**3.8**	**2.5**	**
Comparison	3.9			5.0			4.2			3.6			**3.4**			**5.0**			3.2			4.5		

café visitors deemed that the human moderator was understanding them better, was easier to understand and was more helpful than the robot moderator and also rated the interaction with the robot as inferior to that with the human moderator (score=3.6). New visitors did also feel that the human moderator understood them better, but reported no significant differences understanding the human or robot moderator or in how well the robot or human moderator helped them. They also rated the two interactions as equally good (score=5.0).

Comparing between different experienced and first time language café visitors, the statistical analyses with Welch two samples t-test showed that there were no significant differences between the two subject groups for: feeling understood by the moderator (regardless of if it was a robot or a human), understanding the moderator, or being helped by the moderator.

The difference between the two subject groups in comparing the interactions with the robot or the human moderator was significant (Chi-square test, $p < 0.05$).

5.1.3. Wizard experience

The human moderator (and hence wizard for the robot) was more familiar with the wizard interface in 5 of the sessions (being the second author in 4 and a colleague in 1), was a returning recruited subject moderator in 2 of the sessions and was a first time recruited subject wizard in the remaining 4 sessions.

With a confederate wizard, subjects reported that the human moderator was significantly better, at understanding them ($p < 0.05$), for understanding ($p < 0.05$) and helping them ($p < 0.001$) than the robot. With a recruited wizard, the subjects found that the human moderator was better at understanding them ($p < 0.01$) and helping them ($p < 0.01$), but that the robot moderator was easier to understand (non-significant difference).

Regarding how subjects felt understood by or understanding the robot, there was no significant difference between wizard categories. The confederates were rated significantly easier ($p < 0.05$) to understand. In both scenarios (human or robot), the difference between how the users felt the moderator could help them with respect to the wizard categories was not significant, even though the confederate (human or robot) was rated higher.

The score when comparing the interaction with the robot to that with the human is higher for the recruited moderators (score=4.5 *vs.* 3.2), and ratings were similar in the robot moderator setting (except the non-significant difference that confederate-controlled robot was perceived as more helpful). A tentative conclusion is hence that wizard experience did not influence how the subjects rated the interaction with the robot.

5.1.4. Features to be improved

The questionnaire suggested a list of features that may need improvement for a successful language café interaction: speech understanding, speech synthesis, interaction style, response time and choice of topics. Participants also had the possibility of reporting features to be improved that were not in the list provided. These are discussed below, together with our observations of the video recordings.

Table 2 shows that almost all of the participants thought that the response time should be improved, and in addition a majority that the related interaction style also needed improvement. The term interaction style leaves room for interpretation, but our understanding of this is the fact that the robot addressed one subject at a time and then turned to the other subject with the same or a similar question, rather than engaging with both of them at the same time.

Table 2: *Ratio of participants stating that feature needs improvement.*

Response time	Interaction style	Understanding of speech	Speech synthesis	Choice of topics
81.8%	54.5%	45.5%	40.9%	36.4%

5.2. Observations from recordings and user comments

From the analysis of the video recordings, the following qualitative observations can be made.

Interaction

The set-up with the robot generally entices more pairwise interaction between one learner and the robot, rather than between peers. This is a caveat for more advanced learners, but may be appreciated by beginner or elementary level learners.

Quite naturally, the learners' L2 level influenced the interaction to a simpler interaction with learners at lower level. Some beginner subjects in fact responded that they preferred the interaction with the robot, since it was more structured.

Interaction between peers increased with language café experience.

The order of the sessions with robot or human moderator influenced the interaction. When the robot session was first the average score for comparing both interaction was also higher (4.5 vs 3.4), although the difference was not significant according to the Welch two-samples t-test performed.

There was a variety of features that participants appreciated in the robot, for instance, the more structured interaction and the fact that he would repeat the question whenever he was unable to understand it.

Interestingly, one participant mentioned that she felt that she ”was not wasting a native speaker's time”, which is something that many people might feel whenever practicing a new language in such scenarios.

Speech technology requirements

Concurring with the questionnaire, observations indicate that the response times for the robot were too long in general, and in particular when the wizard choose to correct an ASR misrecognition of the wizard input. The L2 learners were in fact surprisingly patient when the robot's response time was long (up to several seconds), but the interaction clearly suffered from long response times.

The speech synthesis quality was appreciated for the clarity and simplicity of the robot's utterances, but in the future its speed needs to be dynamically adjustable to facilitate understanding for learners with a lower L2 level and emphasis needs to be specifiable in the synthesized utterances, since this is fundamental for implicit linguistic feedback and clarifications.

Many subjects were annoyed with the robot's filled pause, which was uttered when ASR on the wizard's spoken input was being performed. This was particularly true when the same filled pause was repeated if the wizard had to try again to get the sentence correctly recognised before the robot could speak. Filled pauses should hence be used less frequently and with more variations. Previous studies have shown that Furhat can utilize different multi-modal signals (such as gaze aversion and facial expressions) to effectively claim the floor when responses are delayed [14].

Furhat robot

The robot's automatic ability to turn the head towards the current speaker (c.f. Section 3) was found inconsistent by several subjects according to the questionnaire. Observations further indicate that this may in part have been due to timing conflicts between the automatic turning and the wizard's manual request, so that the head turned in the opposite direction to the wizard's intention. It is apparent that the automatic decision of whom to turn the head towards can not be based only on acoustic level, but needs to take interaction state and modeling of the robot's intentions for the following interaction into account.

Related to the above, subjects also requested improvement in gaze, presumably to get an improved experience of the robot actually looking at the person that it is interacting with.

In some sessions, there were facial animation issues, resulting in that the robot's lips were not moving when speaking or were moving when not speaking. This likely influenced the subjects' perception of the interaction, and in the questionnaire, some participants requested more facial expressions.

A background story for the robot is needed. Learners were returning questions to ask robot about personality, preferences or technical details, based on role play, interest in the technology or to play the system.

A robot personality can improve the interaction by introducing jokes, irony and interaction style, as this increases learner engagement in the role play with the robot. This is also something that subjects commented on as being appreciated.

Wizard-of-Oz

A more structured scripted interaction sequencing is required to improve interaction flow and shorten response times (in particular for less experienced wizards). In the questionnaire, wizards requested more buttons with more sentences, backchannels and the possibility to control the robot's facial movements. When observing the wizards in action, it seemed that they spent time browsing the different tabs and buttons in the interface, unsure of what could be said with pre-generated sentences, thus prolonging response time compared to if fewer alternatives, which were relevant to the current interaction state, had been presented in a state-chart based interface that would change over time.

Compared to being on-site moderator, the wizards were less comfortable with being mediated (shift of 1 on the scale 1–5, judged that they understood the participants almost as well (86% responded 4 or 5), and that the participants understood the robot almost as well (71% answering 4 or 5), but that it was much more difficult to assist the learners with linguistic problems (72% on 1 or 2) and that the most difficult was to respond quickly to direct questions to Furhat.

Wizard experience affected the interaction, but the possibility for spoken wizard input using ASR clearly facilitated the wizard task for inexperienced wizards, and as seen above, subjects did not rate the interactions with the robot controlled by naïve wizards lower. Figure 2 shows that most wizards employed the ASR input option substantially more than pre-generated utterances. It is noteworthy to observe firstly that the two sessions with lower than 50% spontaneous utterances were from sessions when the moderator started in the robot setting and secondly that the two sessions with almost exclusive spontaneous input were for the moderator with the most experience at leading the normal language cafés at KTH.

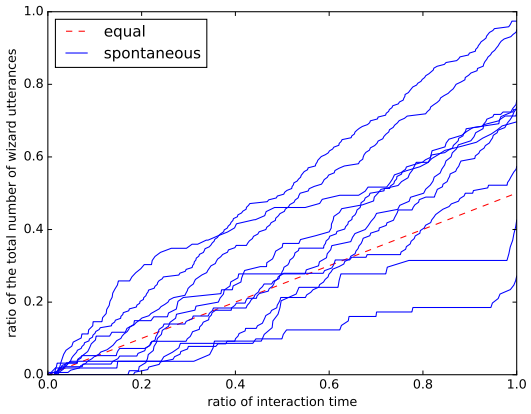

Figure 2: *Incremental ratio of spontaneous wizard utterances for different wizards and sessions. The dashed red line indicates equal use of pre-generated and spontaneous utterances.*

6. Discussion

As expected, the interaction with the human moderator was significantly better than the one with the robot moderator in almost every dimension, except for understanding the moderator, where the difference was non significant and some participants actually felt that the robot was easier to understand. Another interesting fact is that even if we can assume that the understanding by the robot was at human level, since it was controlled by the wizard, there was a significant difference in the extent to which participants felt understood. We attribute this difference mainly to the long response times, and to some extent to inconsistencies in the robot's turning of the head, which may have given the impression that the robot was not understanding.

The results reported in Section 5.1.2 show that in general participants who had never been to a language café before were significantly more enthusiastic about the interaction with the robot than those that usually attend language cafés. This might be due to the fact that the latter may have some expectation about what a language café should be like and they were disappointed with the interaction with the robot. This is something that should be taken into account, since it seems that the target audience was not entirely satisfied with the interaction.

Most results related to different factors in Section 5 were not significant, partly due to the fact that the number of subjects in each category is low and several factors are varied simultaneously. However, as a qualitative interpretation, we conclude that the intermediate speakers probably had a too high level for the robot language café to provide appropriate training, at least in the current state. The self-reported intermediate were in fact rather fluent in Swedish, and even though they made occasional mistakes in grammar or vocabulary, they rarely hesitated or asked for support and help, unlike the subjects at beginner and elementary level, who on the other hand appreciated the simpler, structured practice. Rather than A2 level being the lower threshold, as we initially targeted, we therefore conclude that learners who have reached at least A1 level may be more appropriate for future development of the robot language café, and that above B1 constitutes the upper limit above which interactions with the robot is not fruitful, until the system has reached a state where it is able to engage in more complex discussions.

We would have expected that interactions where a confederate with more experience using the wizard interface would have been smoother than those with a recruited wizard, but the results presented in Section 5.1.3 on the contrary showed that the average score for interaction comparison was higher with a recruited subject as wizard. However, as the confederates were rated higher than the recruited moderators in the human moderator setting, it is unclear whether the difference was that the recruited wizards were judged to perform better or if it was rather the relative difference compared to the human moderator setting that was smaller, because the confederate was also more at ease with being human moderator. More data would be needed to understand how the wizard experience affects the moderator's interaction.

7. Conclusions and Future Work

This paper presents a preliminary analysis of a data collection in a language café setup with a robot. We have described the setup and reported an analysis both from the questionnaires filled out after the interaction and from observations of video data. Despite the wizard-of-oz setup used, which assumes perfect understanding capabilities and language generation, there were significant differences between the participants perception of the human and the robot moderators. The response time from the robot in the wizard setup is the most urgent feature to be improved in the near feature. From the data and also from the observations of the video recordings, the appropriate target level to interact with the robot in a language café scenario should be in the range A1 to B1. Participants with such level will create chances for the robot to incorporate strategies used to help participants such as completing phrases or correcting mistakes automatically. In addition, conversations with this target group will more likely have a more limited range of topics, thus being easier to process from the speech understanding point of view. On the other hand, their pronunciation can be an obstacle for the use of speech recognition. Another challenging feature to implement is handling of the quite commonly used strategy to ask how a given English word is said in L2. This requires the ability to handle code switching between Swedish and English.

In the near future we will improve this setup removing the automatic filled pause before Furhat starts to speak, providing easier access to backchannels filled pauses, and facial expressions. We will further generate a more dynamic interface based on the collected data to train a statistical model that can predict the continuation of conversations within the conversation topics that occurred.

8. Acknowledgements

This work is supported by the Swedish Research Council grant 2016-03698 Collaborative robot-assisted learning of Swedish as a second language.

9. References

[1] S. Eurobarometer, "Europeans and their languages," *European Commission*, 2012.

[2] H. Morton and M. A. Jack, "Scenario-based spoken interaction with virtual agents," *Computer Assisted Language Learning*, vol. 18, no. 3, pp. 171–191, 2005.

[3] P. Wik and A. Hjalmarsson, "Embodied conversational agents in computer assisted language learning," *Speech communication*, vol. 51, no. 10, pp. 1024–1037, 2009.

[4] W. L. Johnson and A. Valente, "Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures." in *AAAI*, 2008, pp. 1632–1639.

[5] H. Hautopp and T. Hanghøj, "Game based language learning for bilingual adults," in *ECGBL2014-8th European Conference on Games Based Learning: ECGBL2014*. Academic Conferences and Publishing International, 2014, p. 191.

[6] A. Neri, C. Cucchiarini, and H. Strik, "Feedback in computer assisted pronunciation training: Technology push or demand pull?" in *in Proceedings of CALL Conference CALL professionals and the future of CALL research*, 2002, pp. 179–188.

[7] J. Han, "Emerging technologies, robot assisted language learning," *Language Learning and Technology*, vol. 16, no. 3, pp. 1–9, 2012.

[8] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: A survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[9] S. Lee, H. Noh, J. Lee, K. Lee, G. Lee, S. Sagong, and M. Kim, "On the effectiveness of robot-assisted language learning," *ReCALL*, vol. 23, no. 1, pp. 25–58, 2013.

[10] O. Mubin, S. Shahid, and C. Bartneck, "Robot assisted language learning through games: A comparison of two case studies," *Austral J of Intelligent Information Proc. Systems*, vol. 13, 2013.

[11] H. Admoni and B. Scassellati, "Roles of robots in socially assistive applications," in *IROS 2014 Workshop on Rehabilitation and Assistive Robotics*.

[12] J. Kennedy, P. Baxter, E. Senft, and T. Belpaeme, "Social robot tutoring for child second language learning," in *The Eleventh ACM/IEEE International Conference on Human Robot Interation (HRI '16)*, pp. 231–238.

[13] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granstrom, *Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction*. Springer.

[14] G. Skantze, M. Johansson, and J. Beskow, "Exploring turn-taking cues in multi-party human-robot discussions about objects," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 67–74.

[15] G. Skantze and S. Al Moubayed, "Iristk: a statechart-based toolkit for multi-party face-to-face interaction," in *Proceedings of ICMI*.