UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Experiment and Evaluation:

A Robot as an English teacher

CIR - Cognitive Interaction with Robots

Students:
Arthur ANDRIEUX
Micha ł BORTKIEWICZ
Ana COCHO BERMEJO
Kevin HUESTIS

Course 2020-21

Abstract

The purpose of this project is to design and develop an engaging robotic english language teaching system. A proof of concept system was built that is used in an experiment to evaluate what makes Human-Robot-Interaction engaging.

The proposed hypothesis is that users prefer to learn language using voice interfaces over keyboard interfaces. Also, two secondary hypotheses will be tested. The first one regarding the direct relationship between a user's age and importance of voice interaction. The second one will be the one testing if making the robot show emotion will offer a more enjoyable experience.

Keywords: *Cognition, HCI, Robot, Language learning, speech recognition, RAL, HRI*

Contents

Abstract **1**

1. State of the art **3**

1.1 Robot-assisted language learning 3

1.2 Social interactions 5

1.3 Human Robot Interaction 6

1.4 Statistical Analysis 7

2. Proof-of-Concept System **7**

2.1 Requirements 8

2.2 User Interface 8

2.3 Speech Recognition 9

2.4 Text Matching 10

3. Experiment **10**

3.1 Participants 11

3.2 Procedure 12

3.3 Questionnaire 13

4. Evaluation **15**

4.1 Hypothesis 1 15

4.2 Hypothesis 2 17

4.3 Hypothesis 3 19

5. Conclusion **19**

6. Bibliography **21**

7. Questionnaire **23**

1. State of the art

1.1 Robot-assisted language learning

Robot-assisted learning is a niche area in social robotics and human-robot interaction (HRI). Advances in speech recognition and in robotics as well as a better understanding of the importance of language skills have helped enormously to this ubiquitous interes in RALL. As Randall et al. [4] state in their recently published evaluation of different methods for *Robot-assisted language learning (RALL)*, RALL is becoming a more commonly studied area of human-robot interac- tion (HRI). Their investigation deepens on theories and methods from many different fields, different instructional methods, robots, and populations to evaluate RALL's efficacy. Also the authors analyse the characteristics of robots used being the main characteristics analysed the following,

- *Function:* autonomous vs. teleoperated
- *Form:* based on their appearance: anthropomorphic, zoomorphic, mechanomorphic, or cartoon-like. Anthropomorphic robots are often considered superior for supporting language learning. When recommending desired characteristics, Chang et al. [8] include the importance of a human-like appearance, which is postulated to increase student engagement and attribution of the robot as a real conversational partner.
- *Voice,* being a synthetic voice or a pre-recorded being the use of synthetic voices most common in RALL they are produced by text-to-speech (TTS) software allowing different accents, etc. Pre-recorded voices are advantageous, as stated in [10] in that they more readily carry emotion but on the other hand, poor quality output may result in decreased language comprehension skills.
- *Social Role as* teacher, teacher's assistant, peer/tutor or learner.
- *Verbal and Non-verbal Immediacy*[1]
- *Non-verbal cues.* Overall, there is a tendency to use facial expressions and body movements to convey action, intention, and emotion [4]. Used in RALL are head nods and happy faces, smiling, thumbs up gesturing, head shakes, sad faces or LED color changes corresponding to incorrect or correct answers; eye gaze for attention guiding, turn- taking, and emotional display; sound effects and music for emotional emphasis or to increase energy, color and light changes to show emotion, etc...
- *Personalization,* using an individual's name, customization of content to one's ability and feedback customization based on one's progress.

  When analysing the results, if splitting the language according to Chen et al. "from subword, word, clause and up to discourse/text " RALL states that robots have been so far used to influence learning at all levels being this research unbalanced:

[1] Non-verbal immediacy is defined in [4] as creating a feeling of connectedness with others by smiling, gesturing, leaning in, engaging in eye contact, using a pleasant vocal tone, reducing physical distance, and the like.

As stated in [4] the ability of robots to teach *vocabulary* is well documented in RALL being a more interesting question is:

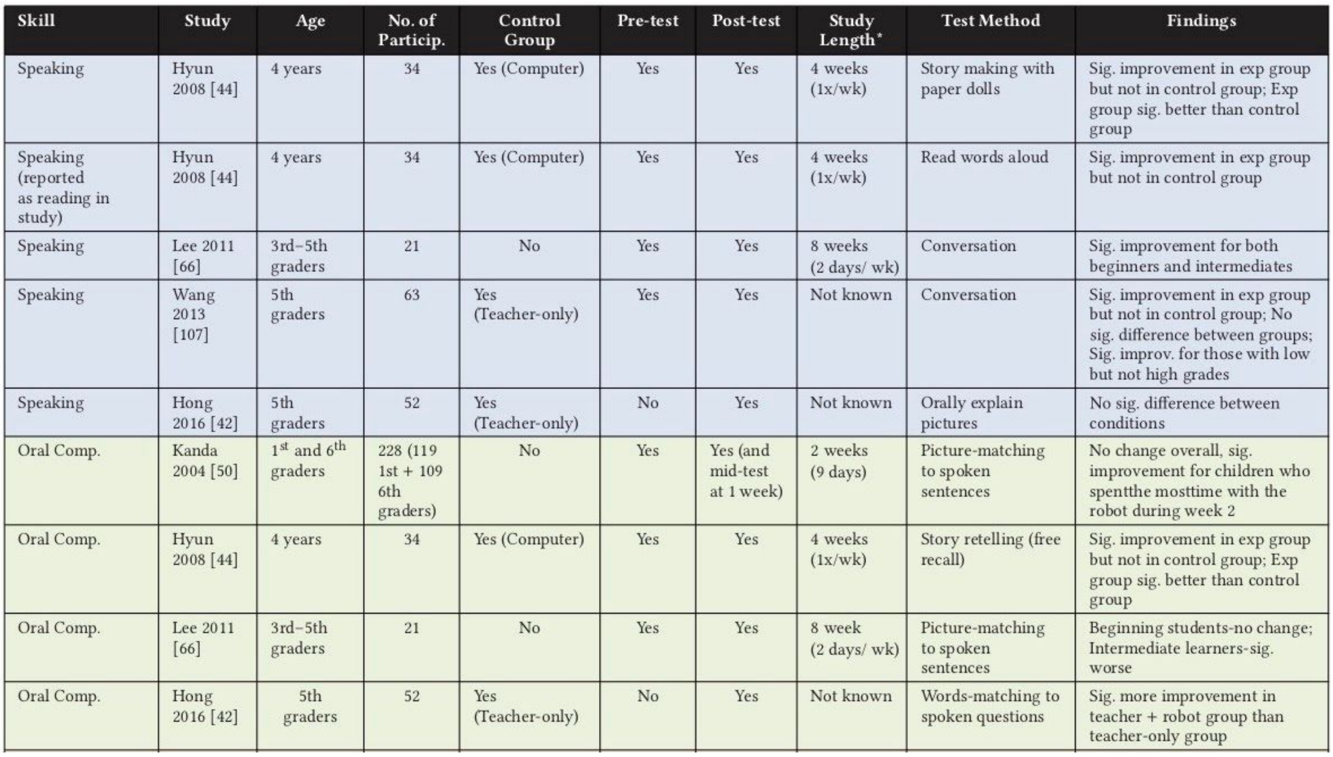
❏ How do robots compare with other technologies in their ability to do so?

Studies suggest that, at least for simple vocabulary teaching, robots perform on par with iPads [15], and for that matter, human teachers.

The influence of robots on *grammar* is not well studied. In the one study to explicitly report these effects, robots significantly improved children's grammar ability as tested after eight weeks of study [14]. Also few studies in RALL have directly measured *pronunciation* improvement. Lee et al. and Wang et al [16] found children's pronunciation significant improvements.

Employing learning- by-teaching paradigm, making the children responsible for teaching the robot Jacq et al. examined how the NAO robot could be used as an accompaniment to a tablet to improve children's writing ability [13].

Although it is likely that robots can help improve reading comprehension , results comparing robot-assistance to other technology is lacking and inconsistent. On the contrary, improvements in speaking are consistently reported while the effects of robots on oral comprehension are mixed.



Skill	Study	Age	No. of Particip.	Control Group	Pre-test	Post-test	Study Length*	Test Method	Findings
Speaking	Hyun 2008 [44]	4 years	34	Yes (Computer)	Yes	Yes	4 weeks (1x/wk)	Story making with paper dolls	Sig. improvement in exp group but not in control group; Exp group sig. better than control group
Speaking (reported as reading in study)	Hyun 2008 [44]	4 years	34	Yes (Computer)	Yes	Yes	4 weeks (1x/wk)	Read words aloud	Sig. improvement in exp group but not in control group
Speaking	Lee 2011 [66]	3rd–5th graders	21	No	Yes	Yes	8 weeks (2 days/ wk)	Conversation	Sig. improvement for both beginners and intermediates
Speaking	Wang 2013 [107]	5th graders	63	Yes (Teacher-only)	Yes	Yes	Not known	Conversation	Sig. improvement in exp group but not in control group; No sig. difference between groups; Sig. improv. for those with low but not high grades
Speaking	Hong 2016 [42]	5th graders	52	Yes (Teacher-only)	No	Yes	Not known	Orally explain pictures	No sig. difference between conditions
Oral Comp.	Kanda 2004 [50]	1st and 6th graders	228 (119 1st + 109 6th graders)	No	Yes	Yes (and mid-test at 1 week)	2 weeks (9 days)	Picture-matching to spoken sentences	No change overall, sig. improvement for children who spentthe mosttime with the robot during week 2
Oral Comp.	Hyun 2008 [44]	4 years	34	Yes (Computer)	Yes	Yes	4 weeks (1x/wk)	Story retelling (free recall)	Sig. improvement in exp group but not in control group; Exp group sig. better than control group
Oral Comp.	Lee 2011 [66]	3rd–5th graders	21	No	Yes	Yes	8 week (2 days/ wk)	Picture-matching to spoken sentences	Beginning students-no change; Intermediate learners-sig. worse
Oral Comp.	Hong 2016 [42]	5th graders	52	Yes (Teacher-only)	No	Yes	Not known	Words-matching to spoken questions	Sig. more improvement in teacher + robot group than teacher-only group

Figure 1. Randall's findings about Speaking & Oral Comprehension [4]

Randall's main findings, on RALL research studies between 2004 and 2017[4], can be summarized as follows:

1. robots are able to help individuals of all ages learn language.
2. robots do not seem to be able to teach vocabulary more effectively than other technology.

3. whether robots are superior to other technology in teaching other aspects of language is still largely an open question.
4. current research suggests that they may offer advantages when used to foster speaking ability.
5. robots seem to aid learning when used as an accompaniment to human instruction.
6. Robots have a positive effect on language learners' affective states: motivation to learn, anxiety when using the language, engagement in the task, and confidence when speaking (both when compared to other technology and when used in tandem with human instruction).

"As stated, when comparing people's talking perceptions with a robot versus a human, people ranked their conversations with people as superior in nearly every regard . This had a lot to do with the human's capability to give appropriate feedback and guidance. The one exception was the evaluation of listening comprehension, where, especially beginners, felt the robot was easier to understand." [11]

1.2 Social interactions

Educational applications sometimes are developed to be able to provide social cues. It is supposed that ability will facilitate natural interaction and so that learning efficacy. Nevertheless as Herberg et al. state in [5] there can be potential costs to social interactions that could run counter to such goals. As so they tested the impact of a watchful versus non-watchful robot tutor on children's language-learning effort and performance. They point out that a better performance on the worksheets was shown in the session in which the robot looked away from, as compared to the session it looked toward the child when filling the sheets.

Also, Breazeal et at. [6] tested on preschool children sociable robotic learning through a robotic companion to support children's early language development. They evaluated whether a robot that "leveled" its stories to match the child's current abilities would lead to greater learning concluding that, despite all children learning new words, children who played with a matched robot used more words and were more diverse.

In fact, physical embodiment[2] has been demonstrated as truly useful for learning interactions with humans. Jung et al. as far as 2004 [7] developed a series of experiments to investigate the relative effectiveness of physical embodiment on social presence of social robots. Experiments showed positive effects of Physical Embodiment of Social Robots (PESR) in the feeling of social presence, the general evaluation of social robots, the assessment of public opinion of social robots, and the evaluation of interaction with social robots. Also, results showed that PESR without touch-input capability causes negative effects.

As so, advantages of using robots [4] can be:

[2] Embodiment as "that which establishes a basis for structural coupling by creating the potential for mutual perturbation between system and environment."

- *intelligence-endowmen*t with sensors that allow communication and interaction with the en- vironment;
- *anthropomorphism-communication* verbal and non-verbal cues can be easily interpreted.
- *accessibility-ability* to be used by people who may not normally have access to instructors.
- *versatility*
- *customization* personalized to both the individual and task.
- *updatability*
- *repeatability* handling repetitive tasks without fatigue.

1.3 Human Robot Interaction

Speaker differences that may result in variability in automatic speech recognition (ASR) include the age, gender, speech rate, and regional dialects of speakers [82]. Language learners produce varied and unique errors that those who have fully acquired the language do not, and this amplifies speech recognition issues. Errors may be syntactic, semantic, or morphological. There may also be phonetic (pronunciation and accent) and pragmatic (contextual and real-world usage) differences that arise. There are still quite a ways to go to make ASR robust for use by language learners in naturalistic contexts, but this step is needed if robots are most useful in the conversational aspects of language learning.

Three prominent theories of language learning but these strategies are best combined in practical language instruction [12]. *Behavioral theory* suggests that positive and negative reinforcement are needed for language learning, along with corrective feedback. *Innatist theory* suggests the need for comprehensible input and an environment that reduces anxiety as much as possible. I*nteractionist theory* states language is best learned when interacting with social others and with the goal of social interaction as so the use of robots is connected to this theory.

As Randall’s concussions shows, gaps in Knowledge in RALL can be stated to be the following:

- *Are robots best employed in a particular area(s) of language learning, and if so, where are they best suited?*
- *How should robots used for language learning purposes look?*
- *Does more learning occur with agents when people prefer the form of the robot?*
- *What age,gender,and cultural effects should be considered when designing robots’ form?*
- *Is exact robot form more important when there is physical presence compared with video presentation, as far as learning gains are concerned?*
- *Is physical presence more important in learning for children and/or beginners?*
- *What identity should robots’ voice project?*
- *Should they employ voices and if so, how should they handle this?*
- *What social roles should robots exhibit or are they best used as tools in language learning?*
- *Should the robots’ affect and engagement strategy be personalized to the user for greater effectiveness?*

1.4 Statistical Analysis

One of the most popular statistical tests used to evaluate different groups of participants in surveys of questionnaires is ANOVA test. It is highly recommended to use this kind of test in scenarios in which there are three and more groups of participants because this test validates if there are significant differences in means between groups. For the purpose of two groups evaluation in terms of means difference, t-test is a go-to method. Using on-tailed t-test we are able to evaluate whether a particular group answers are in general higher or lower than the other.

For the purpose of our experiment we used only above mentioned techniques for hypothesis testing. Due to the low number of participants ANOVA test could not be conducted properly which we describe in more details in section 4.

2. Proof-of-Concept System

When designing a proof-of-concept system, it was important to firstly, keep the hypotheses in mind (see section 3) and secondly, to have the focus on a human-robot-interaction. For these reasons, the user interface should be well and thoroughly designed. From the hypotheses, it is clear that our robot needs both voice user input methods and keyboard user input methods. This can be seen in the top-center of the software system overview found in figure 2. Moreover, to make the interaction more engaging and human-like a voice synthesis module should be developed as well.

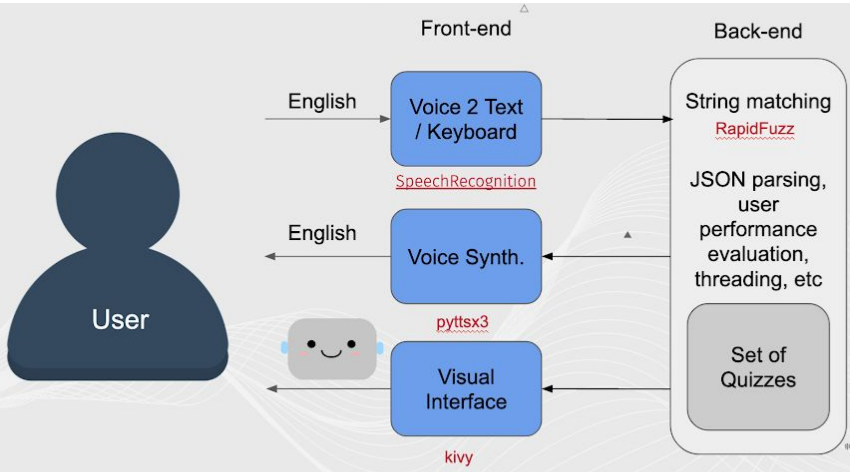


Figure 2. Software System Overview

It should be noted that we decided to develop a system that could be flexibly used on many different devices, from PCs to smartphones, or physical robots that have a touch screen. In the following, we limit ourselves to only PCs and laptops, but the code could be run on e.g. a raspberry pi or smartphone with a touchscreen without any modifications. This setup also provides the chance of future development of the system to be included in a physical robot. More on this in sections 2.1 and 2.2.

A user interface should provide a simple way for users to interact with these modules, and provide additional functionality. This will be further discussed in section 2.2. Due to reasons presented in section 2.3, a string matching algorithm is needed, before the user's input can be processed by the backend.

The backend of the system does not warrant much discussion. The quizzes, as well as the text's that the robot reads as an introduction, transitions, etc. are saved as json files. These are parsed, using common json parsing. The participants' scores are not saved as an output, but rather we rely on a user self evaluation regarding their english level in the questionnaire (see annex).

2.1 Requirements

There are two sets of requirements for the system: development and compilation; and launching the application. Since we compile the application to create executable files, these executables can be run without installing the necessary libraries for development. It is only necessary to have OpenGL 2 installed on a Windows or Linux operating system. That is all for the first set of requirements. Since these requirements are minimal, smartphones or raspberry pis running Linux can also run the compiled executables.

Moving on to the requirements for the development or compilation of the code. Firstly, the system was developed fully in python (some statistical analysis was done in Matlab, which is why there are also Matlab files in the repository). For development we used Python 3.7 and above, but it is feasible that lower versions can also run the code. We did not do extensive testing, to determine the minimum versions of python of any of the libraries included. (Although kivy 2.0, an essential package used, requires at least Python 3.6). We also used Anaconda (anaconda.org) to facilitate the development, however, this is not a strict requirement.

As mentioned, the system is made up of many modules (e.g. voice synthesis and string matching). Many of these modules did not have to be developed, instead relying on python packages that met our requirements. A list of packages used (except common packages such as NumPy, PyYAML, etc.) and the modules that the correspond to is:

● User Interface:	Kivy==2.0.0
● Speech Synthesis:	pyttsx3==2.90
● Speech Synthesis:	espeak
● Speech Recognition:	SpeechRecognition==3.8.1
● Speech Recognition:	pyaudio
● String Matching:	rapidfuzz==0.14.2

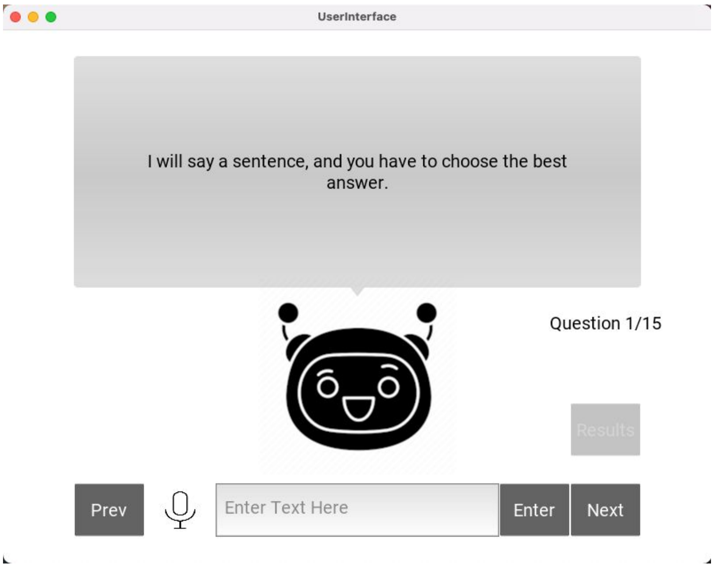
It should be noted that the code currently does not run on Mac computers, because there is a bug in the speech synthesis library pyttsx3 when using it in combination with threading in python. Surely, this can be fixed, but instead we focused our development on Windows and Linux instead. This did not have a great impact on the number of participants, since most Mac users managed to get access to a windows PC as well.

2.2 User Interface

The main package important to the user interface is Kivy. It is described on the website, kivy.org, as being an "open source Python library for rapid development of applications that make use of innovative user interfaces, such as multi-touch apps.” Again, making the system deployable on almost all imaginable types of devices.

The layout of the interface can be seen in figure 3 below. All of the buttons (“Prev”, “Enter”, “Next”, Microphone) can be pushed at any time. However, because of our implementation, using prioritized threading, the previous threads might be completed first. For example, when pressing “Next” the speaking thread is finished, but all other threads in the queue are deleted and the interface moves to the next question.

Pressing the microphone icon (which is hidden and disabled for the keyboard group), the user can speak, giving a voice input. The input is only collected after the robot announces “I’m listening”. The keyboard input can be given by writing in the text field and either clicking on the enter button or pressing the enter key on the keyboard.


UserInterface

I will say a sentence, and you have to choose the best answer.

Question 1/15

Figure 3. User interface of the group that has the option of choosing between both input methods (keyboard and voice)

All speech by the robot is written into the speech bubble above the robot image. The exception is the question asked. Since the quizzes are oral comprehension, we hide the question asked and rely only on the speech synthesis. The speaking is fairly clear, so that no participants complained (except above the speed of speaking).

After asking the question, the possible answers are listed in the speech bubble and the robot waits for the user input. Once the user has input their answer, the evaluation is done and the robot gives it’s response. In the case of a correct answer, the robot announces this and automatically moves on to the next question, while continuing to smile. If the response was incorrect, the robot announces this, reads the correct answer and frowns.

This is a very basic way to show emotions and should be improved in future systems, e.g. by animations and intelligent algorithms, determining the exact emotion that should be expressed.

2.3 Speech Recognition

One of the possible user interfaces is a speech interface available in group one and three. For this purpose we incorporated speech recognition using the SpeechRecognition library in Python that allowed us to use several state-of-the-art voice recognition APIs. We used the default Google Web Speech API which is based on recurrent neural network architecture composed of LSTM cells. Using this kind of out-of-the-box technology we were able to focus mainly on the interaction between core components rather than spent hours implementing our solutions for every issue.

For the purpose of voice recognition we implemented the SpeechRecognizer class located in speech2text.py file that is responsible for configuring the microphone interface and consists of a speech recognition method called recognize_speech_from_mic. As an output from this method we receive a dictionary that consists of information whether there occurred any error and if not we get the transcription in form of a string. It is worth mentioning that due to imperfect neural models and noises occurring during the recording there is a low probability of receiving a perfect transcription. Therefore we had to incorporate additional measures to properly carry on the voice to string matching.

2.4 Text Matching

Due to the incorrectness of speech transcription we had to implement some metric of similarity between the transcription and actual answers located in our json file to guess the answer said by the user. For this reason we implemented additional string processing using RapidFuzz library that calculates the similarity score between two strings based on the Levenshtein distance. Levenshtein distance is basically the minimum number of single-character edits (insertions, deletions or substitutions) required to change one sentence into the other. It turned out that the combination of this metric and basic speech recognition carried on with Google Web Speech API was enough to correctly guess the actual answer which the participant spoke. The process of string matching is presented on Figure 4.

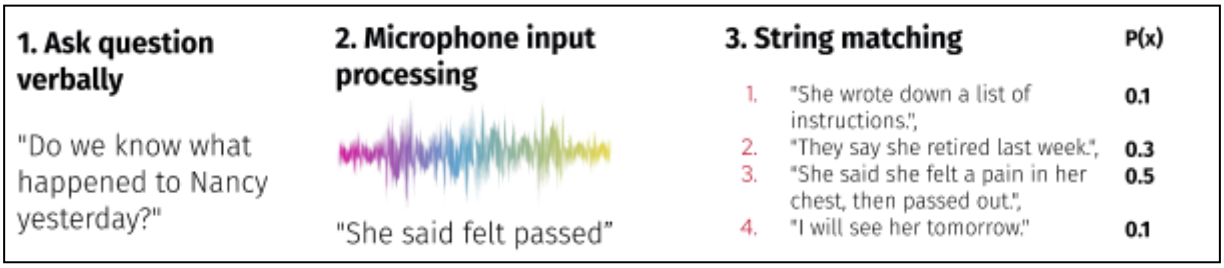

1. Ask question verbally
"Do we know what happened to Nancy yesterday?"
2. Microphone input processing
"She said felt passed"
3. String matching
P(x)
1. "She wrote down a list of instructions.", 0.1
2. "They say she retired last week.", 0.3
3. "She said she felt a pain in her chest, then passed out.", 0.5
4. "I will see her tomorrow." 0.1

Figure 4. Three steps of answer guessing based on Speech Recognition and RapidFuzz libraries. After asking the question user input is processed to obtain the raw transcription. Next the raw transcription is compared with the answers from the database using Levenshtein distance. Eventually the closest answer from the database is chosen as an answer of the user.

3. Experiment

Our project aimed to use the robot to test hypotheses concerning interactions between humans and robots, more specifically how can robots and artificial intelligence improve tasks that already exist. The hypotheses we tested were the following:

H1. Do users prefer to learn a language using voice interfaces over keyboard interfaces?

H2. The older the user is, the more important vocal interaction is when using technology.

H3. If the robot shows emotion, the experience becomes more enjoyable.

Our primary focus was on H1, as we deemed it was the most important. The other two appeared as secondary objectives that we could assess to support our main goal. To confirm or refuse or hypotheses we designed an experiment that we will describe in the following.

3.1 Participants

We contacted 48 participants aged from 15 to 65. To balance our experiment we had to remove a few participants from our list so that the 15-24 years old age bracket was not over represented. Some participants could not participate because of technical issues or unavailability during our testing period.

After this we had a group of 29 participants that completed the experiment and gave their feedback. They were all people with a level in English that spanned from a minimum knowledge of the language to full fluency from native speakers.

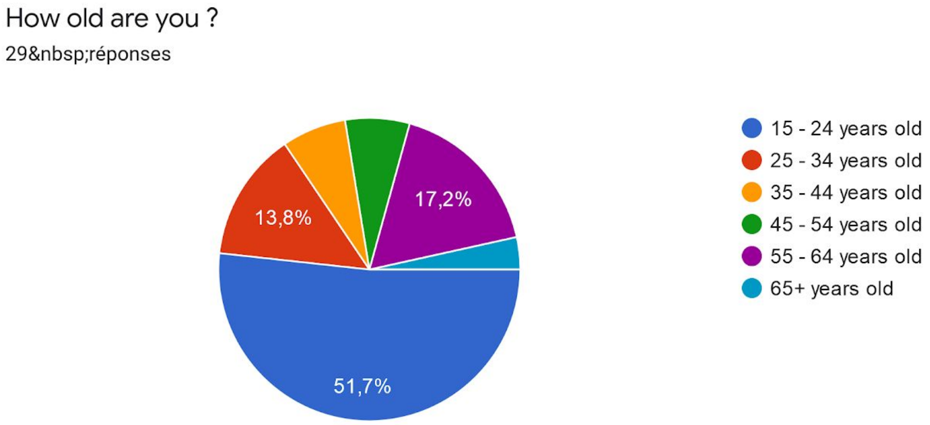


Figure 5. Age distribution of participants

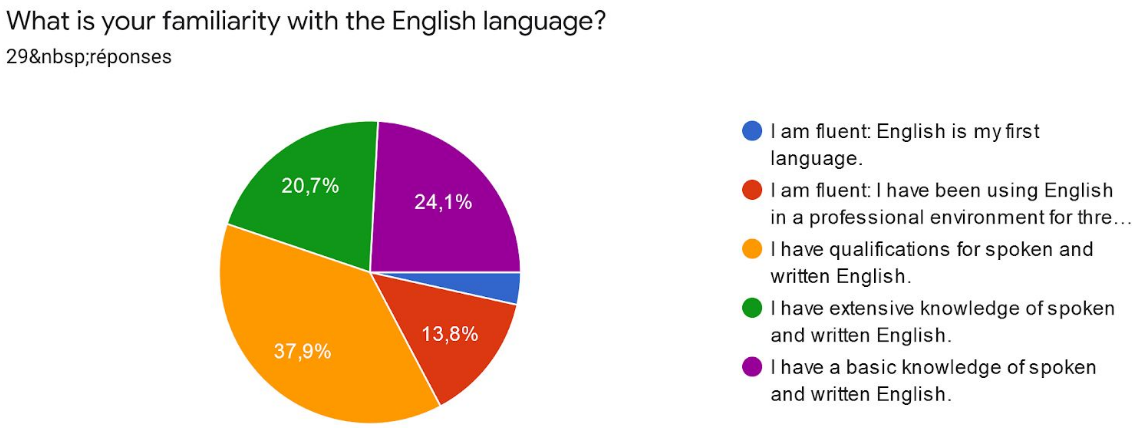


Figure 6. Participants level of fluency in English

We divided the participants in three groups to better evaluate our project. We distributed the groups in terms of age and English knowledge so that we did not introduce any bias. The groups were the following:

- The “control” group: the feedback from the user is given using only a keyboard.
- The “vocal” group: the feedback from the user is given using only the voice of the user.
- The “mixed” group: the feedback from the user is given using either voice or keyboard, at the participant's discretion.

For each group, there was a different version of the app (a different executable was compiled) that presented only the relevant features e.g., the control group could not answer using a microphone. This was done by, e.g. disabling and hiding the text box for the vocal group.

3.2 Procedure

The experiment was presented as a test of the participant’s skills in English. To accomplish this in an efficient and credible fashion, we based the experiment on English tests for certifications in a professional setting such as the TOEIC®, TOEFL® or IELTS™.

In the experiment, each participant individually answers questions testing the participant’s skills in English. The experiment begins with a few introductory sentences to present the robot and the conditions of the test to the participant. The robot then says a sentence without displaying on the screen. Then, a list of four different possible answers are displayed, and the user has to choose the one they think is the most appropriate and either type or say it fully, depending on which group they are in.

The questions and introduction are said by the robot through a text-to-speech interface and at the same time, an animation representing the robot is displayed on the screen.

- If the correct answer is given, the animation displays a smile and the user is congratulated.

- If the wrong answer is given, the animation displays a frowny face and the user is given the correct answer.

After this, the robot moves on to the next question. There are a total of 15 questions in the test.

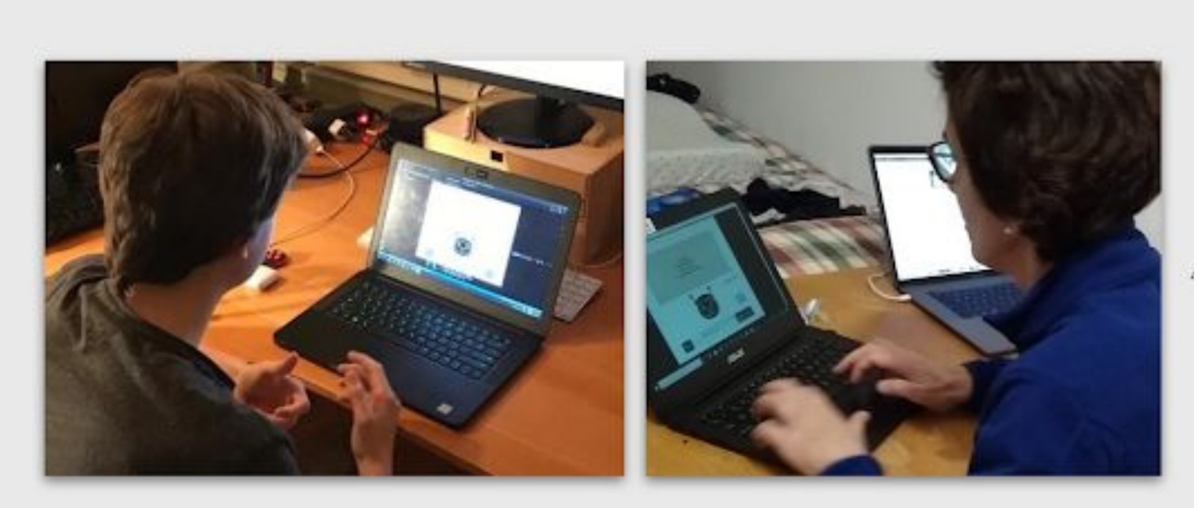


Figure 7. Participants from two different age groups using the system.

Certification tests we mentioned above evaluate the level of the participant based on four major skills: oral expression, oral comprehension, written expression, and written comprehension. These skills can easily be evaluated by our robot with a few minor alterations, but in this experiment we choose to focus on testing only oral comprehension (understanding the question asked by the robot) and oral expression to a lesser degree (speaking well enough so that the robot understands you).

To keep the experiment short and entertaining for the participants, the test needs to be short yet complete. The test will be split between oral and written comprehension, with questions on each skill placed at random. There will be 10 oral questions and 10 written questions.

Once the 15 questions are done, the user has the choice of reading their results, and according to how well they did, the robot displays either a happy or frowny face again. The user is then given a link to a Google Form questionnaire to evaluate their experience with the robot.

3.3 Questionnaire

The survey is inspired by the one developed in the papers by Elaine Short and al [1] and Bainbridge et al [2]. We choose those papers as inspiration since the objectives fall in the same domain as ours, evaluating how the robot interaction can provide a better experience for the end user. The questions were divided into three categories.

First of all we had 4 questions about the participant themselves:

- What group were you in? (1, 2, or 3, the participants did not know that there was three different versions of the app)
- How old are you?
    - 15-24 years old
    - 25-34 years old
    - 35-44 years old
    - 45-54 years old
    - 55-64 years old
    - 65 years old and more
- What is your familiarity with the English language?
    - I am fluent: English is my first language.
    - I am fluent: I have been using English in a professional environment for three years or more.
    - I have qualifications for spoken and written English.
    - I have extensive knowledge of spoken and written English.
    - I have a basic knowledge of spoken and written English.
- Have ever you used any language learning app such as Memrise or Duolingo? (Yes - No)

Then we had 11 questions to help us confirm or refuse our hypotheses:

- How much did you enjoy the experiment? (Between 1 and 7, meaning respectively “Not at all” and “Very much”)
- How much did you enjoy the interaction with the robot? (Between 1 and 7, meaning respectively “Not at all” and “Very much”)
- Do you think using this kind of device would help you learn English better? (Between 1 and 7, meaning respectively “Not at all” and “Very much”)
- Would you use this kind of device in the future? (Between 1 and 7, meaning respectively “Very unlikely” and “Very likely”)
- Would you suggest using this kind of language learning app to people over 60? (Between 1 and 7, meaning respectively “Not at all” and “Very much”)
- Would you suggest using this kind of language learning app to people below 10? (Between 1 and 7, meaning respectively “Not at all” and “Very much”)
- Have you experienced any malfunctions? (Yes - No)
- Hearing the robot speak to me and ask me questions was more interesting than if I would have just read the questions. (Between 1 and 7, meaning respectively “Very unlikely” and “Very likely”)
- I think that facial expression of the robot made the interaction more engaging. (Between 1 and 7, meaning respectively “Very unlikely” and “Very likely”)
- I think that talking to the app or robot while learning a new language is more efficient than just keyboard interaction. (Between 1 and 7, meaning respectively “Very unlikely” and “Very likely”)
- If you were in group 3, what was your preferred input method? (Voice - Keyboard)

And finally three free answer questions to evaluate the system itself:

- Please describe any malfunctions that occurred.
- What would you add to the interaction to make the learning experience more enjoyable?
- Additional comments.

The full questionnaire is attached as an annex. Using the results from the questionnaire, we proceeded to evaluate our system and the hypotheses.

4. Evaluation

Due to deployment issues we could not carry on the experiment on the intended number of participants. However, eventually we were able to distribute the windows version of our application to achieve at least the necessary number of participants to perform the evaluation. In the following subsections we include analysis performed on 28 participants.

We decided to stick to the one-tailed t-test for means for the purpose of evaluation of hypothesis 1 between group one (voice input) and group two (keyboard input). As a backup plan for this hypothesis we also included an additional question addressed to the group three (mixed input) which directly asked the participants whether they were using voice or keyboard interface more eagerly. Moreover we asked every participant whether they prefer talking to the app rather than writing responses.

Initially one-way ANOVA was considered for trying to determine if means of different age groups were significantly different, especially in hypothesis 2, as several groups of data were to be compared. However, due to the low number of participants older than 24 years old we could not perform a valid statistical test using this method. The results we received using ANOVA test were completely insignificant from a statistical point of view. Therefore, we mainly just describe graphically and verbally insights found during the data analysis.

Evaluation of the optional hypothesis 3 was impossible to perform because of not interactive enough application. Most of the participants did not realize the changing facial expression of the robot based on the correctness of given answers, probably because of the small size of the robot face.

4.1 Hypothesis 1

> “Do users prefer to learn a language using voice interfaces over keyboard interfaces?”

As mentioned previously, we performed a one-tailed t-test for means for two most important questions related to the first hypothesis on group one and two. These two questions were:

1. 'How much did you enjoy the experiment?'
2. 'How much did you enjoy the interaction with the robot?'

Both of these questions were to be answered on the scale from 1 to 7, where 1 means completely not enjoyed and 7 completely enjoyed. On Figure 8, we present box plots of answers to these two questions for every group.

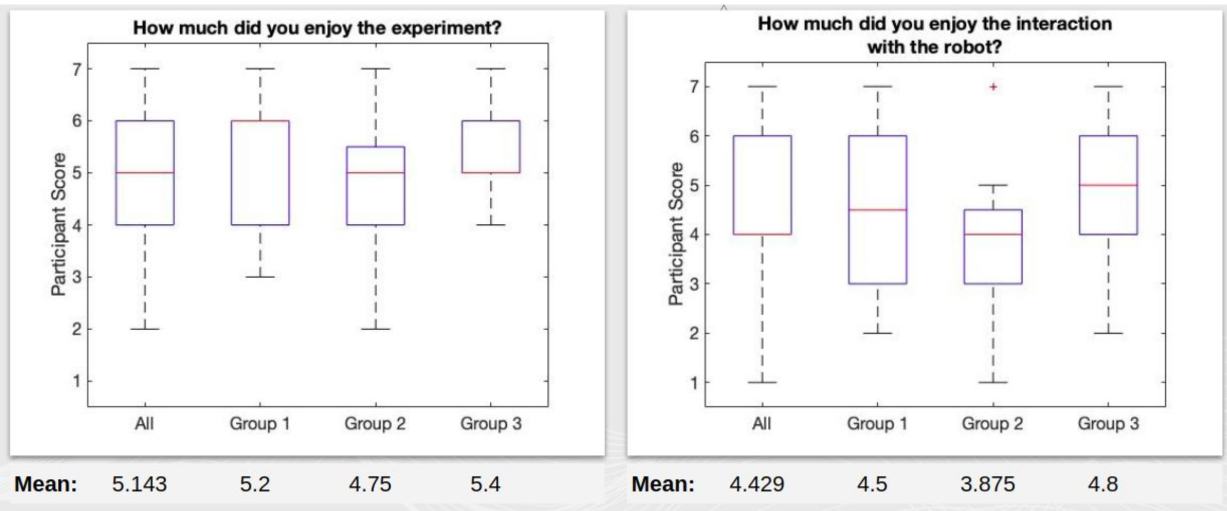


Figure 8. Boxplots of answers given to two mainly related questions to hypothesis 1

The t-test scores and p-values are presented in Table 1. As indicated by the p values there is no strong evidence for rejecting the null hypothesis. However, perhaps by interviewing a larger number of participants we would decrease the p value to a standard threshold of around 0.05 because there is a visible tendency of higher enjoyment of the experiment in the group 1.

	t value	p value
How much did you enjoy the experiment?	0.660	0.260
How much did you enjoy the interaction with the robot?	0.735	0.236

Table 1. t- and p-values of the two main questions regarding hypothesis 1

To cross-validate our assumptions we looked closely at answers of two additional questions:

1. 'If you were in group 3, what was your preferred input method?'
2. 'I think that talking to the app or robot while learning a new language is more efficient than just keyboard interaction.'

The pie plot of answers to the first of the mentioned questions of the group 3 participants is presented on figure 9. It is clear that in general, when participants had the possibility to choose whether they wanted to use keyboard or voice interface, they rather chose the voice one.

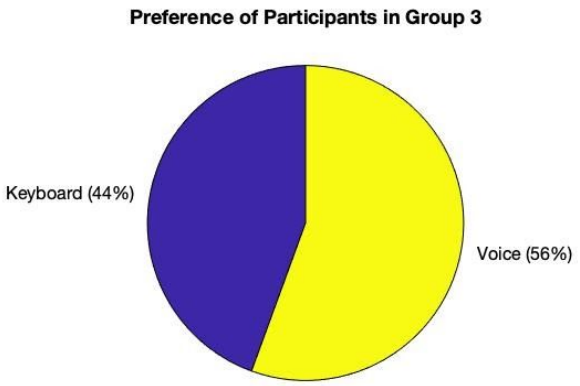
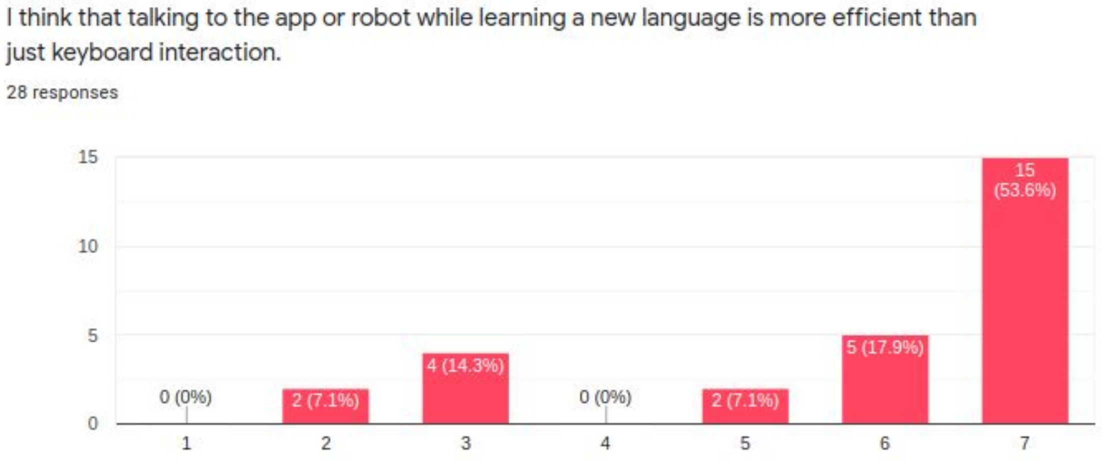


Figure 9. Pie chart of preference of interface (Group 3).

The histogram of the answers to the second question is presented on figure 10. We can see that the majority of participants voted on 7 out of 7 which means that they completely agree with the statement. Moreover, 22 out of 28 answers were in general positive which indicates the correctness of our hypothesis.



I think that talking to the app or robot while learning a new language is more efficient than just keyboard interaction.

28 responses

Figure 10. Histogram of answers given to above question.

4.2 Hypothesis 2

“The older the user is, the more important vocal interaction is when using technology.”

Particularly focused on validating hypothesis 2 were two questions on the questionnaire: *“Would you suggest using this kind of technology for people below 10?” and “... over 60?*

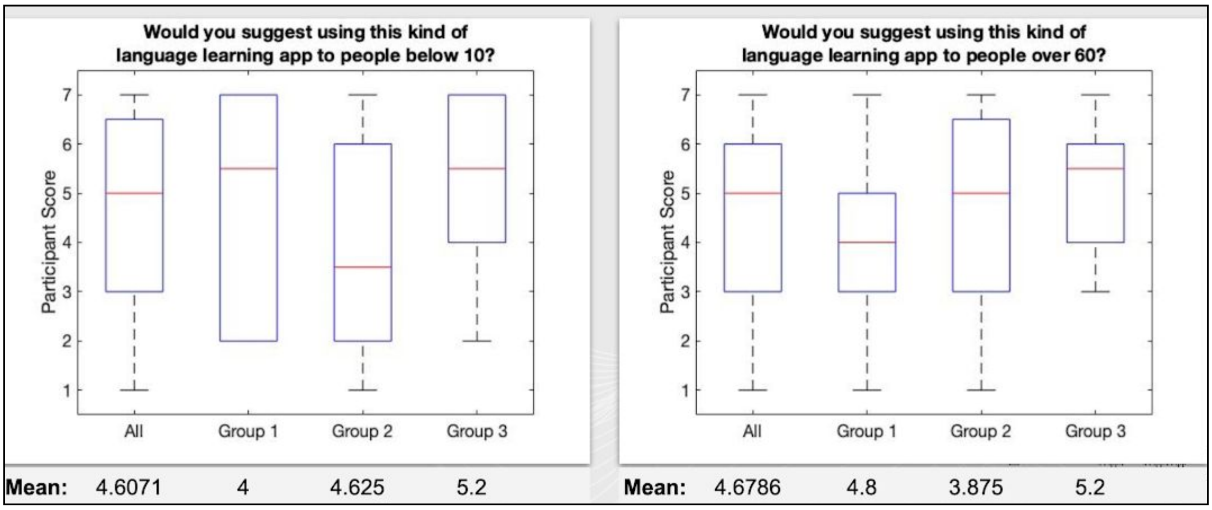


Figure 11. Box plots of answers to questions related to recommended age of the app user.

As we can understand from the data plots and analysis general mean about recommending this technology was pretty similar, around 4.6 over 7. Also group 1, (voice) recommended in a higher way the technology to kids not recommending it so intensely the group 2(keyboard group). This situation can imply the idea that kids are not yet good at writing will find the keyboard writing input not adequate.

On the other hand, when talking about age over 60 recommendation gets higher for writing. Regarding the other question on which the hypothesis is based :*"I think that talking to the app or robot while learning a new language is more efficient than just keyboard interaction"* it was developed the analysis shown on the figure below.

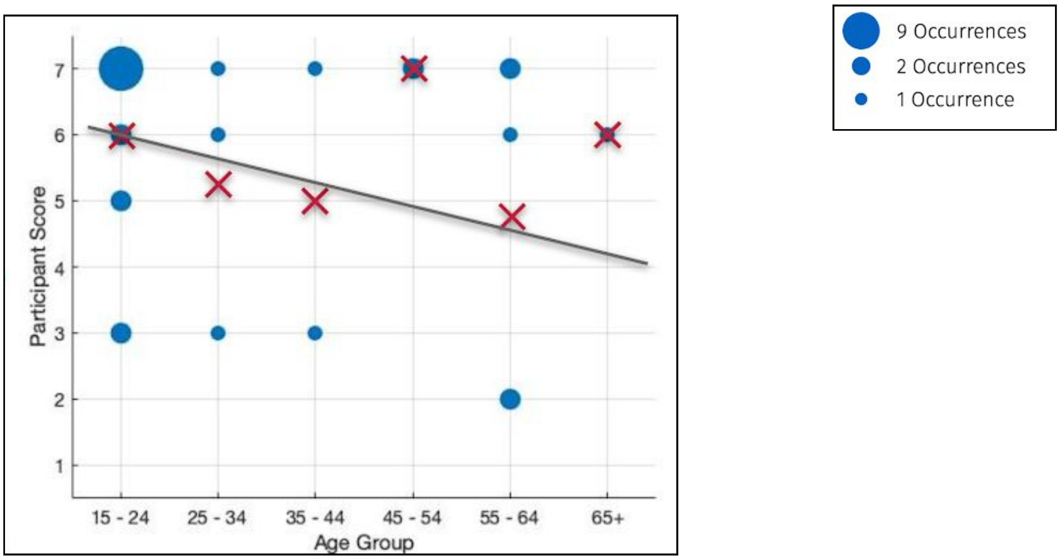


Figure 12. Answers given by the participants to the question "I think that talking to the app or robot while learning a new language is more efficient than just keyboard interaction" based on the age of the participant. Size of the bubble represents the number of users, the line represents the probable tendency in the data.

As we can understand the plot it is clear that engagement through this technology for language learning is starting to decrease over 35. Nevertheless there was not enough data to evaluate answers for this question for ages over 65. Also answers between 45-55 were most confident on the answer (but this can imply the group was pretty biased in relation to its knowledge of technology).

4.3 Hypothesis 3

“If the robot shows emotion, the experience becomes more enjoyable.”

Hypothesis 3 information for analysis was basically based on one question of the survey: “Do you think that the facial expression of the robot made your experience more engaging?.” As we can understand when diagramming the data the idea obtained a general pass for the whole participants as well as for group 1(voice) and group 2 (keyboard) being the confidence of group 3 a little lower.

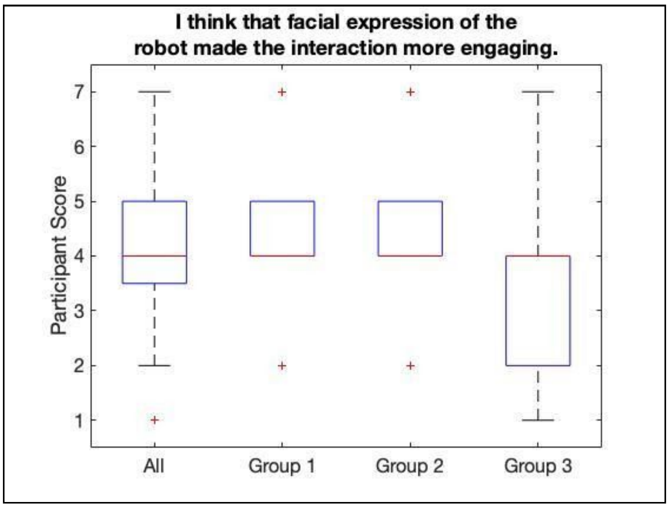


Figure 13. Boxplot of the answers related to the question about robot facial expression.

Nevertheless many participants commented they did not notice the change of emotion on the robot’s face when giving the wrong or the correct answer. As so, no other conclusion can be taken regarding the hypothesis due to the lack of data.

5. Conclusion

To conclude, we developed a working proof-of-concept system that can easily be used for further development and used with a wide range of devices. Using this system, compile as an executable file for Windows and Linux, we conducted an experiment with 28 participants.

The experiments focused on answering one main hypothesis: “Do users prefer to learn a language using voice interfaces over keyboard interfaces?” Using the t-test method, we were not able to reject the null hypothesis, however, the questionnaires filled out by all participants showed fairly strong evidence to support the hypothesis.

The secondary hypotheses were explored along similar lines. However, since the majority of participants were between the ages of 15 and 24, we did not have enough participants to prove or disprove the second hypothesis. Interestingly, the data tended towards evidence against our initial hypothesis. This might be due to the implementation of the voice interface and issues with the technology though. The third hypothesis could not be further tested, since the implementation of “showing emotions” was rather weak, such that some participants remarked that they had not noticed that the robot shows emotion at all.

In total, as a group we learned very much about cognitive interaction with robots and the experiments that come along with its development. We learned that software development (i.e. threading and the user interface) can be very challenging, in some instances taking up as much time as the development of the robot’s intelligence. Also, making sure that the exported system runs smoothly across a wide range of applications can be very challenging.

Regarding the experiments, it was reinforced that Murphy’s Law that “Anything that can go wrong will go wrong” is particularly relevant in these sorts of experiments. This means that the participant instructions should be very clear and precise.

Finally, from the user’s short responses in the questionnaire, we suggest two possible future improvements to the system. These remarks were: “Questions too hard for inexperienced English users” and “Robot should speak slower”. This could be addressed by giving the robot more intelligence. For example by adjusting speaking speed and question difficulty automatically using reinforcement learning algorithms.

6. Bibliography

1. Short, E., Hart, J., Vu, M. and Scassellati, B., 2010, March. *No fair!! an interaction with a cheating robot.* In 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 219-226). IEEE.
2. Bainbridge, W.A., Hart, J., Kim, E.S. and Scassellati, B., 2008, August. *The effect of presence on human-robot interaction*. In ROMAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication (pp. 701-706). IEEE.
3. Williams, L.J. and Abdi, H., 2010. *Fisher's least significant difference (LSD) test. Encyclopedia of research design*, *218*, pp.840-853.
4. Randall, N. 2020. *A Survey of Robot-Assisted Language Learning (RALL).* ACM transactions on human-robot interaction. Vol. 9, nº1-7. DOI: 10.1145/3345506.
5. Herberg, J. Feller, S. Yengin, I. Saerbeck, M. 2015. Robot watchfulness hinders learning performance. In Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (ROMAN'15).
6. Westlund, J.K. and Breazeal, C., 2015, March. The interplay of robot language level with children's language learning during storytelling. In Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction extended abstracts (pp. 65-66).
7. Jung, Y. and Lee, K.M., 2004. Effects of physical embodiment on social presence of social robots. Proceedings of PRESENCE, pp.80-87.
8. Chang, C.W., Lee, J.H., Chao, P.Y., Wang, C.Y. and Chen, G.D., 2010. Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school. Journal of Educational Technology & Society, 13(2), pp.13-24.
9. Gordon, G. Breazeal, C. Engel, S. 2015. Can children catch curiosity from a social robot? In Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction.
10. Westlund, J. Gordon, G. Spaulding, S. Lee, J. Plummer, L. Martinez, M. Das, M. Breazeal, C. 2015. Learning a second language with a socially assistive robot. In The 1st International Conference on Social Robots in Therapy and Education. Almere, The Netherlands.
11. Lopes, J. Engwall, O. and Skantze, G 2017. A first visit to the robot language café. In Proceedings of the 7th ISCA Workshop on Speech and Language Technology in Education. 7–12.
12. Norbahira Mohamad Nor .Radzuwan Ab Rashid. 2018. A review of theoretical perspectives on language learning and acquisition. Kasetsart J. Soc. Sci. 39, 1 (2018), 161–167.
13. Alexis D. Jacq, Séverin Lemaignan, S. Garcia, F. Dillenbourg, P. Paiva, A. 2016. Building successful long child-robot interactions in a learning context. In Proceedings of the 11th ACM/IEEE International Conference on Human Robot Interaction. 239–246.
14. Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, Gary Geunbae Lee, Seongdae Sagong, and Munsang Kim. 2011. On the effectiveness of robot-assisted language learning. ReCALL 23, 1: 25–58.

15. Westlund, J.Dickens, L. Jeong, S. Harris, P. DeSteno, D. Breazeal, C. 2015. A comparison of children learning new words from robots, tablets, and people. In Proceedings of the Conference Proceed- ings New Friends 2015.

English Teaching Robot

Please respond to the following questions based on your experience with the English teaching robot.

* Required

1. What is your E-mail?

2. What group were you in? *

Mark only one oval.

○ Group number 1

○ Group number 2

○ Group number 3

3. How old are you ? *

Mark only one oval.

○ 15 - 24 years old

○ 25 - 34 years old

○ 35 - 44 years old

○ 45 - 54 years old

○ 55 - 64 years old

○ 65+ years old

4. What is your familiarity with the English language? *

Mark only one oval.

○ I am fluent: English is my first language.

○ I am fluent: I have been using English in a professional environment for three years or more.

○ I have qualifications for spoken and written English.

○ I have extensive knowledge of spoken and written English.

○ I have a basic knowledge of spoken and written English.

5. Have ever you used any language learning app such as Memrise or Duolingo? *

Mark only one oval.

○ Yes

○ No

6. How much did you enjoy the experiment? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not at all	○	○	○	○	○	○	○	Very much

7. How much did you enjoy the interaction with the robot? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not at all	○	○	○	○	○	○	○	Very much

8. Do you think using this kind of device would help you learn English better? *

Mark only only one oval.

	1	2	3	4	5	6	7	
Not at all	○	○	○	○	○	○	○	Very much

9. Would you use this kind of device in the future? *

Mark only one oval.

	1	2	3	4	5	6	7	
Very unlikely	○	○	○	○	○	○	○	Very likely

10. Would you suggest using this kind of language learning app to people over 60? *

Mark only one oval.

	1	2	3	4	5	6	7	
Very unlikely	○	○	○	○	○	○	○	Very likely

11. Would you suggest using this kind of language learning app to people below 10? *

Mark only one oval.

	1	2	3	4	5	6	7	
Very unlikely	○	○	○	○	○	○	○	Very likely

12. Have you experienced any malfunctions? *

Mark only one oval.

○ Yes

○ No

13. Please describe any malfunctions that occurred.

____

Please rate how much do you agree with the following sentences:

14. Hearing the robot speak to me and ask me questions was more interesting than if I would have just read the questions. *

Mark only one oval.

	1	2	3	4	5	6	7	
Very unlikely	○	○	○	○	○	○	○	Very likely

15. I think that facial expression of the robot made the interaction more engaging. *

Mark only one oval.

	1	2	3	4	5	6	7	
Very unlikely	○	○	○	○	○	○	○	Very likely

16. I think that talking to the app or robot while learning a new language is more efficient than just keyboard interaction. *

Mark only one oval.

	1	2	3	4	5	6	7	
Very unlikely	○	○	○	○	○	○	○	Very likely

17. If you were in group 3, what was your preferred input method?

Mark only one oval.

○ Voice

○ Keyboard

18. What would you add to the interaction to make the learning experience more enjoyable?

19. If you have any additional comments, please let us know here. Thank you very much for yo participation !