A physical robot’s effect on vocabulary learning

ANDREAS WEDENBORN
awed@kth.se
2015-10-11

Master’s Thesis at CSC
Supervisor at KTH: Olov Engwall
Supervisor at Furhat Robotics: Preben Wik
Examiner: Joakim Gustafson

TRITA xxxxxxxxx
Abstract
This thesis investigates the effect on having a physical robot present to take the role of a teacher or exercise partner in a language learning exercise. In order to investigate this, an application was developed enabling a vocabulary exercise to take place in three different situations. In one situation a non-embodied voice would be the teacher, in the second situation an animated face on the screen in front of you would be the teacher, and in the final situation the teacher was move into the physical world by a robotic head with a 3D face mask. Initially one study on a vocabulary exercise with 15 participants was conducted. In order to gather more valuable data, the exercise was redesigned and conducted with another 11 participants. The results from the user studies are interpreted to point towards a higher task motivation when working with a robotic head.
Sammanfattning
Contents
Abstract .......................................................................................................................... 2
Sammanfattning ............................................................................................................... 3
Chapter 1 ........................................................................................................................ 4
Introduction ..................................................................................................................... 4
1.1 Scientific question .................................................................................................. 5
1.2 Outline of the report ......................................................................................... 5
Chapter 2 ....................................................................................................................... 6
Theoretical Background .............................................................................................. 6
2.1 The language learning process .......................................................................... 6
  2.1.1 Automaticity ..................................................................................................... 6
  2.1.2 Second language vocabulary learning ......................................................... 7
  2.1.3 Motivation in language learning ..................................................................... 8
2.2 Computer assisted language learning ................................................................ 9
  2.2.1 Benefits of computer assisted language learning ....................................... 9
  2.2.2 Phases and different applications of computers in language learning .......... 9
  2.2.3 Behavioristic CALL ....................................................................................... 10
  2.2.4 Communicative CALL .................................................................................... 10
  2.2.5 Interactive CALL ............................................................................................ 11
  2.2.6 Designing with CALL in mind ....................................................................... 11
2.3 Related work ....................................................................................................... 12
  2.3.1 Virtual language teacher ARTUR .............................................................. 12
  2.3.2 Ville the virtual language teacher ................................................................ 13
  2.3.3 Baldi and Bao ............................................................................................... 14
  2.3.4 Robot tutors ................................................................................................. 15
  2.3.5 Furhat the robotic head ............................................................................... 16
2.4 Summary .............................................................................................................. 17
Chapter 3 ..................................................................................................................... 19
The exercises ............................................................................................................... 19
3.1 The first exercise .............................................................................................. 19
  3.1.1 Language ....................................................................................................... 20
Chapter 7 ................................................................................................................................. 45
7.1 Conclusion ....................................................................................................................... 45
7.2 Future work .................................................................................................................... 45
References .............................................................................................................................. 47
Appendices ............................................................................................................................. 50
Appendix A ............................................................................................................................. 51
List of words used .................................................................................................................. 51
Appendix B ............................................................................................................................. 52
Definition of gestures .......................................................................................................... 52
Appendix C ............................................................................................................................. 53
Questions in the questionnaire ............................................................................................ 53
Appendix D ............................................................................................................................. 55
Test scores ............................................................................................................................... 55
Appendix E ............................................................................................................................. 57
Appendix F ............................................................................................................................. 58
Appendix G ............................................................................................................................. 60
Chapter 1
Introduction

Learning a second language is not a trivial task. On the contrary, to fully learn a new language you must not only understand new grammatical rules but you must greatly expand your vocabulary in order to comprehend the language in different situations. This means taking on a wide array of new words in the second language. The more words you know, the more you will be able to understand of what you hear and read and the more you will be able to communicate. There are many methods of learning new vocabulary for a foreign language, whether it is self studies or studies together with people. One method in language learning and training is through the use of technology and computers, computer assisted language learning (CALL). The subject of CALL has been described as “the search for and study of applications of the computer in language teaching and learning” [1]. It is an ever changing subject as the applications of computers and the technological advancements of computers is constantly changing. Using computers enables a higher degree of freedom since a system has the possibility of being used when it suits the user; it is not dependant on anyone else. Studies by Wik and Hjalmarsson [2] have shown that the attitude towards embodied conversational agents as a complement to a human teacher is positive. Embodied agents are graphically embodied agents who through the use of facial expressions, gestures and speech enable a human-computer interaction that to some extent resembles a face-to-face communication. In the study the virtual language teacher Ville¹ and the role-playing and dialogue system called DEAL², based on Ville, were evaluated with positive results in a virtual learning environment. More about Ville can be found in section 2.3, related work.

With technological advancements it is now possible to move the embodied agent out from the 2D screen to an actual 3D surface. Using a physical robotic head with an animated projection onto a 3D mask, the face can move an embodied agent from the 2D screen to the reality of 3D. Previous work regarding physical robots in language learning is much limited to looking at the people’s attitude towards using physical robots in the learning process, not what the actual results could be. This report will focus on what effects can be measured and what effects are perceived from moving the virtual teacher from the screen to an actual 3D robotic head in a language learning environment, more specifically a vocabulary learning environment.

¹ http://www.speech.kth.se/ville/
² http://www.speech.kth.se/deal/
1.1 Scientific question

How is the language learning process of vocabulary learning affected by the physical presence of a robotic head with a 3D face?

This question will be evaluated by constructing a vocabulary training exercise followed by a test of the words in the exercise. This exercise will be performed in three different settings – with a non-embodied voice as a teacher, with an animated face on screen as the teacher and finally with a robotic head with a 3D projection as the teacher. This effect will be evaluated based on results in the test after the exercise and evaluated by a questionnaire following the exercises.

1.2 Outline of the report

In chapter 2 a theoretical background related to this thesis is presented along with a review of related work. Following that, in chapter 3, the methodology and implementation is explained. Chapter 4 describes the user study and chapter 5 moves onto the actual results of the user studies. Finally chapter 6 discusses the results and chapter 7 summaries the work as well as gives thoughts on future work.
Chapter 2

Theoretical Background

This chapter gives a short background to the skill acquisition in the grammar and vocabulary learning process, followed by how computers have been used in computer assisted learning environments and a summary of related work.

2.1 The language learning process

2.1.1 Automaticity

“When learning a second language, what we really want to achieve is automaticity”, Wik concludes in his doctoral dissertation [3].

Automaticity is when a task can be completed with no effort. The person does not have to think through the process step by step in order to complete the task, but can perform it while thinking about other things. A person speaking in their native language is an example of this, the task of conveying a message is done with no specific thoughts about actually forming functional sentences, whether it is regarding the correct grammatical construction or the actual physical movement of the tongue and mouth to shape the words. When learning a second language, this is initially what one would struggle with, as there are a lot of new aspects to a language. But at the same time this is the ultimate goal when taking on a new language. Speaking the new language without step-by-step translation from the native language to the new language, and without carefully thinking about pronunciation, is moving towards achieving automaticity in the new language.

Wik further describes Anderson’s adaptive control of thought (ACT-R) [4]. There is a distinction between declarative and procedural knowledge. Declarative knowledge is simply the fact and information a person knows, know-what knowledge. Whereas procedural knowledge is skills a person knows how to perform, know-how knowledge. In the declarative stage, instructions are encoded in the brain as a set of facts. These facts are actively used when performing a certain task. As the task is practiced more and more, a person develops a procedure for the task and the active access to the declarative access becomes less and less needed. Anderson describes three stages of skill acquisition; the three stages are summarized as follows by Wik.

- The Cognitive stage: In the first stage the learner receives instructions about a skill in declarative form. Explicit information is given in order to provide clear and concise rules and sufficient examples, which the learner can interpret and rehearse, thereby raising awareness
of and internalizing the skill. The processing in this stage is conscious, deliberate, slow and requires full attention.

- **Associative stage:** The major development of this stage is knowledge compilation. During the associative stage, a process of proceduralization takes place, converting declarative facts into production form. The learner should here be offered opportunities for abundant repetition within a narrow context (which is what drills are all about).

- **Autonomous stage:** After a skill has been compiled into a task-specific procedure, the learning process involves an improvement in the search for the right production. In this stage, the procedure becomes more and more automated and rapid. The process underlying this stage is tuning. Three learning mechanism serve as the basis of tuning: generalization, discrimination, and strengthening.

Studies on grammar and vocabulary acquisition have shown that these stages also apply to the process of developing automaticity in a second language [5] [6].

These theories form the very basis of vocabulary learning, vocabulary can be regarded as the smallest piece of instruction. The task of forming a sentence involves perhaps several steps in order to get the correct word in the correct place. But vocabulary learning can be broken down to just one instruction, the actual translation of the word. And in vocabulary learning we also want the process of translating of a word to achieve automaticity. Though often vocabulary learning is not as simple as just the word, it is expanded with relevant information regarding the morphology of the word. But the very beginning of vocabulary can be considered to have the instruction of translation which we want to move to an autonomous stage.

### 2.1.2 Second language vocabulary learning

Oxford and Scarcella discusses the process of vocabulary learning and methods of learning new vocabulary [7]. They describe many factors which affect vocabulary acquisition. These factors include maturational constraints, frequency and attention.

Maturational constraints discuss the affects of maturity and how it affects the ability to learn. While some aspects of learning a second language are more difficult for adults, Oxford and Scarcella conclude that maturational constraints do not prevent adults from learning new vocabulary; we are constantly learning new words to our vocabulary throughout our lives. Unlike children who are not conceptually able to learn some new words, adults can comprehend the meaning of more words. For
instance young children have a difficult time of comprehending the meaning of before and after, and empty and full, while adults have no problem understanding their meaning. This enables the author to not pay too much attention to the age when designing an exercise.

The aspect of frequency relates to vocabulary learning through reading text. Arguments have been made [8] where people are more likely to learn words that appear more frequently in a text when it is read. This, however, does not relate to the exercise in this thesis as it will not involve any reading of a longer text.

A very interesting point which Scarcella and Oxford discusses is the effect attention has on the vocabulary learning process. Studies [9] promote the fact that the amount of attention the learner gives to a word determines how well it will be remembered. Interesting ways of learning a new word could grab the learner’s attention more, thus perhaps resulting in increased learning. Emotions can also directly affect attention [8], if the learner is put in an emotional state where he is excited and interested it can have an increased effect, where if the person is bored it can have the opposite effect. Other influences like the words meaning to the person, and the word itself may contribute to how easy they are to remember.

2.1.3 Motivation in language learning

In order to learn something as big as a new language, one most likely requires some sort of motivation. Generally motivation can be categorized into two different classes of motivation - intrinsic and extrinsic [10]. Intrinsic motivation is driven by the enjoyment or interest in the task itself, the pleasure of learning or feeling what they are learning is important. Extrinsic motivation is based on rewards and incentives outside the actual task. For example if the task is completed successfully, the person would be given an amount of money. Extrinsic motivation can also relate to feelings, such as the anticipation of punishment, for example bad grades on exams. [11]

Motivation can also come from the task itself, task motivation [3]. The task motivation seems to be additive, in other words it is more motivating to do a task that is fun and enjoyable than a task that is not, with all other motivational factors being equal. Thus, one could benefit from having a fun and enjoyable task but still having the same intrinsic and extrinsic motivation regarding it.

Finding the correct motivation is perhaps one of the most important features to keep a person interested in the continued learning.
2.2 Computer assisted language learning

2.2.1 Benefits of computer assisted language learning

There could be many benefits to computer assisted language learning, CALL. A computer will never grow tired of repeating learning tasks for when it takes longer for a person to learn something new. An exercise could be run over and over again, with the computer providing the same valuable feedback with no sign of exhaustion after repeated runs. It could also present the learner with a greater influence over how the method or exercise should be formed, making it easier for the user to tailor their experience to increase their learning.

Isaacs and Senge writes about [12] how group dynamics such as peer pressure and any anti-learning pressures could be avoided with computer assisted learning. In a teacher situation, a student might be afraid to ask a question regarding something while successfully showing understanding within something else. The teacher could get the impression that the student understands everything along the way, and the result is gaps of knowledge because the teacher cannot question everyone about everything and the students are afraid to ask. In computer assisted learning, when done in an individual environment, there is no pressure from the teacher in comparing students, perhaps allowing students to ask questions no matter what. What is being taught and learned is between the student and the teacher, in this case the computer.

In a group environment, such obstacles are harder to overcome as they still exist. Although methods can be applied to continuously encourage positive feedback towards each other and positive group dynamics.

2.2.2 Phases and different applications of computers in language learning

Computers have been used for different kinds of language teaching for over three decades. Warschauer & Healey [13] has summarized the development and history of CALL, computer assisted language learning, where they divide it into three stages; behavioristic CALL, communicative CALL and integrative CALL. The different stages represent different pedagogical approaches and the development of a new phase does not mean a complete rejection of an old phase. Below is a short introduction to the different phases, followed by a discussion why they relate to the work in this thesis.
2.2.3 Behavioristic CALL

The earliest phase of CALL was implemented in 1960s and 1970s [14]. The ideas behind it were based on the current dominant theories of general learning. Theories around this time mainly revolved around repetitive language drills which were referred to as “drill and practice”. This made computers a good choice of a tutor. “Computers serves as a vehicle for delivering instructional materials to the student”, Warschauer describes. In general, it revolves around the following statements;

- Exposure to the same material over and over is beneficial, perhaps even essential to learning.
- The computer does not get bored with carrying out repeated tasks, and it can do so without providing judgmental feedback, thus making it ideal for carrying out repeated tasks in language learning.
- A computer can adapt to its student, presenting material on an individualized level, allowing students to advance at their own pace.

In the 1970s and early 1980s, the theories behind behavioristic CALL faced two problems. It was rejected as both a theoretical and pedagogical level, and personal computers were on the rise. This spurred the development of the next phase.

2.2.4 Communicative CALL

The communicative approach to CALL was moving away from the drill and practice-method used earlier. John Underwood meant that the previous phase did not allow for enough authentic communication to be of value [15]. He developed a series of premises for communicative CALL, which are as follows:

- Focus on the communication rather than on the form, avoid drills
- Teach grammar implicitly rather than explicitly.
- Allow and encourage students to generate original utterances rather than just manipulate prefabricated language.
- Do not judge and evaluate everything the students do, nor reward them with congratulatory messages.
- Avoid telling students they are wrong and be flexible to a variety of student responses.
- Use the target language exclusively and create an environment in which using the target language feels natural, both on and off the screen.
- Never try to do anything that a book can do just as well.
This approach leaves the computer in a more passive role, where perhaps the computer waits until summoned to respond to requests and questions from students[16]. This approach was not exclusive in comparison to the older model; many implementations could still make use of drill and practice-scenarios. However the students were more encouraged to use one computer per group and then discuss their answers within the group before inputting anything in the computer. At times the difference between behavioristic and communicative CALL is mainly how the software is used rather than a major difference in the implementation.

Communicative CALL had many advantages over the previous methods, but many educators felt that it did not live up to its full potential - the computers were used in an ad hoc and disconnected fashion [17]. A number of educators wanted the language learning process to be in a more integrated manner, a more authentic social context, for example using task- or project-based scenarios. Thus a new challenge was presented, and with advancements in technology it would lead to a new phase.

2.2.5 Interactive CALL
Since the beginning of the twenty-first century, advancement in technology has allowed for a whole new creation of software for CALL. Working efficiently with different media, such as pictures, graphics and sounds, different scenarios can be created to provide a scenario much like an authentic social scenario. This typically means creating simulations of real life events, such as ordering at a restaurant or asking someone for directions. These simulations are enhanced by using audio and visual elements in an attempt to submerge the user into an interactive scenario, much like the real world.

With the technological advancements today, all three different phases are implemented in different ways. No method is generally considered to be superior today; they all work in their own ways and the results depends very much on what fits the user the most.

2.2.6 Designing with CALL in mind
At an early stage of planning of the user study for this thesis, the goal was to separate the different exercises depending on what visual aid you had acting as a teacher, or exercise helper. Without any visual aid of an animated face or robot, the intention was to keep the instructions short and informative in text focusing on giving a word and its translation in a clear way with the purpose of having little to no social aspects of the exercise. Any instructions would be text based and no audio
would be played. The exercise with an animated face on the screen were to serve the purpose of having a virtual teacher on screen to help the user feel more engaged with a teacher on screen. The exercise would still be very straight forward, but with a virtual teacher who would say the words out loud, giving the student an opportunity to get a better idea of how the word is supposed to be pronounced. Finally the implementation with the robotic head with a 3D face were to be more social experience, where the robotic head would go through an exercise with more social aspects. The exercise would be narrated and walked through by the robotic head and he would give positive and inspiring feedback during the exercise.

After reading about the three major phases of CALL, it became clear that having such a study would rather put the focus on comparing different methods within CALL. The exercise with the non-embodied voice would perhaps lean towards a more behavioristic CALL method, with the computer helping you in a more drill and practice situation. Involving a physical robot head and tuning the exercise to be of a more social type would lean more towards a more interactive CALL method. The purpose of the study and thesis is to see what effect the presence of a robotic head would have on the language learning process. Therefore to not steer away from the purpose of this, the exercises would have to be designed and held in the same way with the difference being the teacher or leader of the exercise. Either it would be physical in the form of a robot head with a 3D face, present on screen in the form of an animated face or finally through a disembodied voice guiding you through the same exercise. However the exercise itself would be exactly the same, thus making no difference what category of CALL it could represent. This became a crucial part of the designing the implementation and user study.

2.3 Related work

2.3.1 Virtual language teacher ARTUR
ARTUR, the Articulation TUtoR, developed at KTH Speech, music and hearing\(^3\) is a virtual speech tutor which uses three-dimensional animations of the face and internal parts of the mouth, the tongue, jaw and palate. This is key for giving detailed feedback on pronunciation as phonetic features that are hidden in a human speaker can more easily be displayed. Most computer based system for articulation training offer feedback on an acoustic level. For a second language learning and for individuals with hearing impairment this perhaps becomes less relevant as opposed to the physical

\[^3\] https://www.kth.se/csc/forskning/tmh
properties of pronunciation, for example how and where to place your tongue in order to articulate a certain word. Thus having an animated face with detailed features of the internal parts of the mouth can help in the process of learning a new language and its pronunciation. A study [18] on how pronunciation feedback in such software should be given conducted by Engwall and Bälter indicated that the subjects in the study in perceived articulatory instructions to be useful and the paper summarized guidelines on how pronunciation feedback on the phoneme level should be given in computer assisted pronunciation in order to be more effective.

![ARTUR interface showing facial features.](image)

**Fig 2.1: ARTUR interface showing facial features.**

### 2.3.2 Ville the virtual language teacher

Ville is a virtual language tutor developed by Wik [3] at KTH Speech, music and hearing. It is an application and a framework for computer assisted language learning and computer assisted pronunciation training. Ville is an embodied conversational agent that you can talk to and he talks back. He will encourage you, give you feedback and guide you through your development of further language skills. In this software Ville takes on the role of being the teacher. It is currently available as a downloadable application for Swedish vocabulary and pronunciation exercises.

The vocabulary training in Ville is based on ‘Flashcards’. It is a well known vocabulary acquisition tool where normally a deck of physical cards is used. On one side of the card you write the word in your native language and on the other side you write the word in the language you want to learn. The
card can then be used for practicing the translation of the word both ways. This is implemented with virtual cards in the Ville application.

A study [19] with Ville looked at the feedback given from people who used the training program. The study showed that there is a huge demand for verbal training provided outside a classroom environment. The user study consisted of perception exercises, where you listen to an utterance and it is your task to decide what was said (for example ‘bita’ or ‘byta), as well as production exercises, where Ville says the word and you are supposed to repeat it. These exercises moved onto full sentences where single words were changed to full sentences. The response was that perception exercises are much easier to perform. It also showed that creating good feedback mechanics for production exercises is important. In general it could be seen that students who performed more poorly appreciated the language teacher more, which shows that it can be helpful for those most in need of help.

![Fig 2.2: The interface in the Flashcard pronunciation training with Ville.](image)

**2.3.3 Baldi and Bao**

Baldi is a computer animated talking head with features of visible speech synthesis. The project of developing Baldi was led by Ron Cole at the University of Colorado. The big purpose of Baldi is to be of help for individuals with hearing problems who would benefit from the visual help of a face animating what is being said, both to easier see what is being said but also for general training in facial expressions together with speech. Baldi has three visible layers, see Figure 2.3, where you can see the face as a whole, semi-transparent and as a full wire frame.
In 2004 Slim, Cohen, and Massaro worked on extending their current virtual teacher Baldi to be multilingual [20]. The work done was aimed towards extending Baldis feature with another language. Their work is mainly focused on the technical aspects of adding a new language to such software. Another language means another ways of pronunciation, which means other ways of animating the face.

In 2006 Massaro, Dominic W., et al built their own extension to Baldi [21]. This extension focused on creating a Mandarin speaking agent, whom they named Bao. Their work was focused on children with language challenges as well as all individuals interested in learning a second language. As opposed to previously mentioned work with Baldi, the work with Bao involved actual user testing - both among children with hearing difficulties as well as people learning mandarin as their non native language. Their user studies concluded that the use of the visible speech agent did not clearly facilitate pronunciation learning relative to just auditory speech, but the authors believe that a more prolonged training period would show an advantage of visible speech.

2.3.4 Robot tutors
In 2012 the result of a 100 participant study on the role of a robot’s physical presence in a tutoring task was presented [22]. The study consisted of the task of solving a puzzle. The puzzle was a Nonogram⁴. The users were tested on four different puzzles in three different settings. The first setting was with a tutor present as a disembodied voice. When solving a puzzle the tutor would

⁴ https://en.wikipedia.org/wiki/Nonogram
randomly pause the puzzle three times to give a short lesson or tips about nonograms, such as “I have an idea that might help you”. In the second setting the tutor was visualized in the form of a video representation of a robot. The robot on screen would supply similar tips as in the previous scenario, though during this the robot would look towards the user and towards the board when speaking. Finally in the last setting the robot was actually present next to the user. Again, giving the same type of feedback as in the previous cases but this time moved into the physical world as a robot. The users did not test all systems; they only tested one system each.

Their study investigated what effect the physical embodiment of a robot tutor had in a task like this. It showed that participants who received lessons and tips from a physically present robot actually outperformed participants who received the lessons and tips from a disembodied voice and a video representation of the robot. From these results they draw the conclusion that physical embodiment can yield measurable learning gains in robot tutor interactions.

2.3.5 Furhat the robotic head

Skantze et al. have built Furhat [23]. Furhat is a robotic head that consists of a 3D mask with an associated projector. The projector projects an animated face onto the 3D mask creating the visual effect of having a live 3D animated face. The name Furhat comes from the fact that the common way of covering the projector and mechanics is by putting a hat on the head, giving a sophisticated looking robotic head.
The projection of the animated face upon the 3D mask gives the feeling of the robot speaking when the face is animated. Furhat can make use of a Microsoft Kinect to work as its eyes. The facial tracking enabled by Kinect allows the user to be positioned in 3D-space and allows the robot to track the movement of the user and focus its eyes on the user. The Microsoft Kinect together with a stereo microphone connected to the system allows the system to handle multiple users, allowing social interaction in a multiparty environment to take place.

Furhat was developed at KTH and is now part of a company called Furhat Robotics. The ‘brain’ of Furhat is a computer running the framework IrisTK. IrisTK was developed together with Furhat and will be explained further in the implementation chapter.

2.4 Summary

The work in this thesis started with researching the field of computer assisted language learning, language learning in general and virtual and robot tutors. The research has led the author to believe:

- The concept of learning new vocabulary can be affected by such factors as attention, as it puts you in a different emotional state which affects your ability to take in and remember new information.

- There are different ways of using computers in the language learning process. Computers can be the main source of interaction, or it can serve as a passive unit which can be consulted if
help is needed. It can present you with an interactive dialog and situation simulation, or it can serve the purpose of just helping you by playing back the utterance of a word

- Motivation can come in different shapes. One motivation that seems to appear additive to the effect of others is the actual task motivation, how fun and engaging the exercise itself is.

- Virtual language teachers have seen positive feedback in some areas, with the focus on the possibility of giving better and individualized feedback.

- Robot tutors have been seen to increase the performance of puzzle solving when they are physically present rather than just present on the screen.

- Furhat is a robot which with its framework enables social interaction with an animated face and actual head movement.

After doing this research it was decided that in order to avoid having the user study focus on the different aspects of computer assisted language learning, the focus should be on the actual effect of lifting the virtual language teacher from the 2D screen to a 3D physically present robot. Much like the puzzle solving study was able to see some improvement with the physical robot present; it is of interest if the same kind of behavior can be seen within vocabulary learning. As mentioned this led to the conclusion that the exercise to be implemented and tested in a user study should be the same across all platforms, with the difference between them being the interaction of a virtual teacher. One instance of the exercise would have a non-embodied voice, the second instance of the exercise would have an animated agent on screen and the final instance of the exercise would have Furhat the robotic head present. With the availability of Furhat on KTH, through Furhat Robotics, it was decided that using Furhat would be suitable for this user study.
Chapter 3

The exercises

In order to construct this exercise with the three different situations, an application was implemented in the IrisTK framework. This section will first briefly explain the idea behind the first implemented exercise, then it will be explained how the framework IrisTK was used to implement this. Finally there will be an explanation of the second implementation of an exercise, which was formed after the first user study had been done.

3.1 The first exercise

In order to learn new vocabulary, you need to take the smallest piece of instruction – the translation of a word – and try to remember this. There are several ways of being exposed to new words and finding their translations. It was of interest to see the vocabulary learning where the computer takes on the role of exercise leader, or teacher, to fully see if there was any effect of having this teacher present as a robot. To achieve this, the exercise could be pretty straight forward. The user could be introduced to reasonable amount of new words, and then be given a small test to see if they remember the words.

The introduction of the word will be in the foreign language, where the computer can show the user how you are supposed to say it, both visually and through speech. An object and the word in text will be presented on a screen along with the computer saying the word out loud. Then the user himself has the opportunity to say it. Should the pronunciation be good enough, he moves on to the next word. If the pronunciation is not good enough, the computer will repeat the word to give you another chance of hearing it. After all the words have been introduced and spoken, the screen will show all the pictures of the words. The computer will then go through the words in the foreign language, one by one, and it is the user’s task to match the correct word with the correct picture in order to show that a basic understanding of a word has been built up. After initial tests and discussions, the reasonable amount of words to be learned in a short amount of time was set to nine.

In order to benefit from what can be seen in task motivation and perhaps the general effect of attention, the exercise was decided to be constructed in a social way where the user was not just presented with a speech playback of the word and then the picture and text. But the exercise would involve a narrative dialogue from the virtual teacher, introducing the task and then through varied dialog take the user through the words one by one providing feedback along the way. This was to make the exercise into a more natural exercise with a teacher rather than just going through a list of words, perhaps increasing the task motivation and the attention of the user.
In short, the exercise will introduce the user to nine words. He will be guided through the process by a voice. They will be introduced to a new word and actually hear the word being said, see a picture relating to the word on screen and see the word in writing on screen. In the exercises with embodied teachers they will also get the opportunity to see the facial features of the teacher saying the word. The user will then get an opportunity to say the word. In order to move on to the next word, the word must be pronounced correctly. When all words have been introduced, a test where the user matches the word the computer is saying with the correct picture will be given. This is all gone through in a clear flow with the computer talking to the user all the time.

3.1.1 Language
In order to do such an exercise it was decided upon a language to learn new words in. In order to conduct the test without too much difference in previous language knowledge, the Russian language was chosen. The prerequisite for the exercise was no previous knowledge in the Russian language, apart from perhaps the knowledge of how to say “hi” and “cheers”. This would enable a broader user base, as the Russian language is not taught in Swedish schools by default.

3.1.2 Different versions
The user study wishes to study whether the physical presence of a robot has a significant effect. In order to attempt to test this, it was decided that the exercise would be given three times - each with the different teacher situations. This meant the same nine words could not be used over again across the systems, meaning another eighteen words had to be introduced in order to create three exercises on different content. The groups of words were all following a theme, they were animals, colors and house related objects. The list of words used can be found in Appendix A.

In order to not spot any differences in difficulty of the words in a set, the different set of words would be used across all situations, meaning the set of words is not tied to a certain test.

3.2 Implementation in IrisTK
IrisTK\(^6\) is a Java-based framework, developed to help constructing multi-modal dialogue systems. It provides developers with the ability to control an animated face which can either be used as a standalone avatar or used in the projection of the face of Furhat. IrisTK provides an API for developers including designing flow based dialogue systems, controlling an animated face and modules for using speech synthesizers.

\(^6\) http://iristk.net/
3.2.1 IrisFlow

IrisTK provides a state-based framework called IrisFlow to define a flow of interaction with the system. It is much like a finite state machine, but with added flexibility. Different states can be structured hierarchically, giving the opportunity to define more generic event handlers as well as more specific events in sub-states. For example, the generic event of a person leaving the interaction, which you would want to handle in every state, can be defined in one state while letting the sub-states extend the parent state, adopting its event handlers. This works very similar to class extension in Java. In IrisFlow you can also define variables in the flow, which affect the execution of the flow. This gives a powerful tool to trigger different events depending on previous events in the flow.

The states can be considered a series of events that you are currently executing, and you will always be in one particular state. Invoking parameters as `<call>` and `<goto>` enables switching of states, where `call` goes temporarily to the called state and executes the events only to `return` to the previous state. But `goto` moves the flow entirely to the new state. The flow is illustrated in figure 3.1 where you can see the `call` command followed by a `return` to return to the state, but `goto` switches states entirely.

![Fig 3.1: Schematic example of state handling](image)

In the implementation IrisFlow is used to construct the flow of the exercise. The different words are handled as states where you go to a temporary state to ensure the word was said correctly by the user before moving on to the next state (word). Finally when all the words have been handled you go
to the state which represents the test at the end of the exercise. In other words, IrisFlow enabled the construction of an exercise like this to be efficient.

3.2.2 Events in the states

In the states you can handle the different events. Standardized events are implemented to features of the face and speech synthesis. For example invoking an event called action.speech will allow the synthesizer to speak and the face to animate what is being said. Events like these can be reused and standardized. For example having the synthesizer and face say something with the possibility of the user interrupting what is being said is defined over 15 lines of code. This has been defined in a reusable file, allowing the execution of such code to be done with just one line of code. Much like methods with parameters are used in programming, methods of doing a series of events can be implemented. This allows for a normal use of the speech synthesis to be used with just one line:

<dialog:say> Say this text </dialog:say>.

Defining reusable events like this helps make the main flow of execution very clear.

In order to make this execution of events, such as dialogue for the computer to speak, more natural and not static, the possibility of defining several events and choosing one at random was used. For example in the feedback after the user has said the word correctly, you can define the flow to only use one out of several possible events. This can be defined within the random tag as follows:

<random>
    <dialog:say> Impressive! </dialog:say>
    <dialog:say> Wow, that was good. </dialog:say>
    <dialog:say> Great! </dialog:say>
    ....
    ....
</random>

When the flow comes to the random event, the flow chooses one of the alternatives within the random tag to execute before moving on. This was taken advantage of in order to create a variation of responses and a naturally flowing speech from the computer, instead of always using the same sentence such as “The next word is”.

22
In the flow of events in the exercise, a graphical user interface, GUI, is used. This interface is used to display the pictures of the words along with the text. When the computer is speaking and animating the face, a separate window is being constantly updated with the correct information depending on where in the exercise you are. A separate class was created to accommodate the GUI. When arriving at the state of a new word, the GUI has to be updated with the correct information. Along with the execution of the speech where the computer says the word the GUI is updated to show the picture and the text of the current word. The text is written in the Latin alphabet, as the users have no knowledge of the Cyrillic alphabet. Finally, in the test at the end, all pictures are presented in a three by three grid.

(a) The GUI when learning a new word

(b) The GUI during the test

Fig 3.2: The GUI during play through
3.2.1 Controlling the face

You are given the ability to control the facial expressions through parameters in IrisTK. These parameters are for example:

- **BLINK_LEFT** - causes the face to blink with the left eye.
- **SMILE_OPEN** - causes the face to smile.
- **SURPRISE** - gives the face a surprised look.

Combining and invoking several parameters within a certain timeframe can enable you to create various expressions of the face, handled as complete gestures. These gestures can then be used to give the face some actual expressions in different situations.

(a) Defining the gesture smile and wink by combining parameters.

(b) The result of the gesture

Fig 3.2: The animated face in IrisTK

Fig 3.3: Creating a smile and wink gesture.
When defined, these gestures can be called upon as an event in the flow, and thus be used to enhance the experience with the animated face and the projection of it onto Furhat. The gestures give a more living impression of the face. In the exercise gestures were used when giving positive feedback after correctly pronouncing the word. The gestures used aimed to enhance the positive feedback when the user said the word correctly. The gestures used where the following:

- smile – The face smiles with a tiny raise of the eyebrows
- opensmile – the face smiles revealing teeth.
- smilewink – the face smiles and gives a wink with one eye.

No negative animations, such as expressing anger, were used as there is no reason for the computer to be angry at the user. The fully defined gestures above can be found in Appendix B.

### 3.2.2 Speech synthesis

The Text to Speech Synthesis (TTS) system used across all applications of the exercise, no matter which level of visual aid, was CerePros TTS called William. This synthesizer gives a natural and expressive voice in most occasions. William provides a British accent and covers all of the talking in the exercise apart from the Russian.

For the Russian synthesis, in order to match the male voice of the rest of the dialogue the built-in Russian synthesizer in Mac OS X Yosemite called Yuri was used. Recordings of each word were stored locally and when the exercise got to the point of saying the Russian word the audio file was played along with a lip sync of the face. The lip-sync of the face was achieved by writing text for the face to mime using similar movement as required to say the Russian word.

### 3.2.2 Speech recognition, feedback and Wizard of Oz

Due to the limitations and implementations of the available Russian speech recognizer, it would be a task exceeding the limitations of this thesis to build a system where Russian would be recognized and properly corrected on the phoneme level. The recognizer tried would simply either understand the word or not understand the word, working very binary. In order to avoid any technical problems during exercise in relation to the recognizer, and because the goal of this study is to collect results and see what kind of effect having a physical robot present would have, it was decided that the

---

7 [https://www.cereproc.com/](https://www.cereproc.com/)
speech recognition part of the exercise would be controlled through the press of a key on a keyboard. This could enable the Wizard of Oz setting during the user study. A person, The Wizard, would serve the role of deciding whether or not the Russian word was correctly pronounced by the user. This gives the Wizard the possibility of invoking dialogue where the computer asks the user to try the word again after having listened to it again or accepting the pronunciation taking the user through to the next word. The user is unaware of the Wizard’s involvement in deciding whether or not the word was correct, meaning the user does not at any point direct his attention to anything but the computer. After the user has said the word in Russian, the Wizard clicks either the Y button on a keyboard for a correct answer, or the N button to give the user another try.

3.2.3 The Kinect
IrisTK provides framework for connecting a Microsoft Kinect to your system. This enables IrisTK to observe the user in a 3D environment, giving the ability of actually looking at the person sitting in front of Furhat. In the exercise with Furhat, the Kinect was enabled and allowed Furhat to follow the user with his vision at all times during the exercise. The exercises without Furhat had no use of the Kinect, thus the functionality of it was disabled during those exercises.

3.3 Summary of the first exercise implementation
In short, the exercise was implemented as a flowing dialogue, almost as a lesson held by a teacher. In this exercise the acting teacher was either a non-embodied voice, an animated face on the screen or Furhat the robot head. The teacher introduces the exercise, that the user will be introduced to new words in Russian, and that a small test will be held after going through the words. The teacher then introduces one word at the time, giving the user an opportunity to see a picture, see it written in text and then finally the opportunity to correctly say it. After having gone through each word, a small test where the user has to listen to a word the teacher says, and correctly identify the meaning of that word by clicking the correct picture. When the test is done the teacher thanks the user for his or her participation and the exercise is over.

3.4 The second exercise implementation
After completing the first user study with the first exercise, it was decided another exercise was to be implemented. The purpose of this exercise was to have more results to discuss and somewhat draw conclusions from. In order to obtain data that could be compared to the first user study, the exercise was redesigned to train the active vocabulary instead of the passive. The passive vocabulary means listening and understanding the word you heard, which is what the test in the first exercise consisted of - the computer said a word in Russian and the user had to understand the word and click the correct picture. Active vocabulary means being able to recall the translation and then saying the
word yourself; you are actively translating and saying the word from memory. This means the test in the exercise would be based on the computer showing only the picture and the user having to translate it and say it correctly in the Russian language. The second exercise was implemented with the same technical aspects as the first exercise, thus the following sections will only discuss the layout of the exercise, not the technical aspects of it. The second exercise was also performed in the three different settings, non-embodied voice, with a face on screen and finally with Furhat.

3.4.1 The training in the second implementation
In order to achieve more measurable data the training phase of the exercise was redesigned. In the first implementation the computer introduced the different words, one by one. In the second implementation the user was in charge in the training phase. After the initial greeting and introduction from the teacher, the user was presented with the GUI as can be seen in (b) in Figure 3.2. The user was informed he had a maximum of five minutes to practice the words. From this point the user would click a picture which would make the teacher present the Russian translation of the word to the user. The teacher would still present the translation of the word through dialogue in order to make sure the focus was lifted from the GUI to the teacher. The users then practiced the different words until they felt ready to take the test, which they started by saying to the teacher that they were ready to take the test. If the five minutes would pass, the teacher would inform the user that the time for practice was over and the test would follow.

3.4.2 The test in the second implementation
The test in the second exercise was very straightforward and not very unlike the training phase in the first exercise. The teacher would show a picture, see (a) in Figure 3.2, but without the text below. It was then up to the user to say the word in Russian. This test was also implemented in a Wizard of Oz setting. The Wizard was given three options when the user said the word. If the word was said correctly or completely wrong, he could either confirm the correct answer or state that the answer was incorrect. The third option was given in case a word was almost said correctly; it would prompt the user to try the word again more clearly. This option was given to help both the user and the wizard, in case anything was unclear, as well as creating a more realistic environment.
Chapter 4
User study

This chapter goes through the user studies that were conducted in this thesis. The first user study was conducted with the first implementation of the exercise; the second user study was conducted with the second implementation of the exercise. Both user studies were set up in the same way, but they were conducted on different days with different users. In this chapter, there is initially a description of the setup of the user studies and the participants of the studies. Finally there is a discussion around the data collection of the user studies. The results will be presented in chapter 5.

4.1 The setup

The user study was set up in a room at KTH Speech, Music and Hearing. The equipment used was a touch screen placed on the table in front of the user. The touch screen allowed users to click directly on the screen during the test at the exercise. This screen would have the graphical part of the exercise at all times, as well as during the exercise with the animated face it would appear next to the display of words. In front of the touch screen Furhat was placed. The user wore a microphone in order to communicate with the computer during the exercise.

The different set of words were used with the different levels of visual aid, giving an equal amount of exercises being done with X visual aid together with Y set of words. The order in which the different visual aids were used varied; some users would start with the robot, some with the animated face and some with the non-embodied voice.
4.1.1 The users in the first study
A total amount of 15 users participated in this study. This included a session of five minutes with each exercise, making the total amount of time taken to complete the exercises around fifteen minutes.

The users were students at KTH personally approached at campus and asked if they wanted to participate in vocabulary learning exercise. A criterion for participating was having no previous knowledge of the Russian language. None of the users were previously known to the author.

4.1.2 The users in the second study
A total amount of 11 users participated in the second study. This included a session of around five minutes with each exercise, making the total amount taken to complete the exercise around fifteen minutes.

The users who participated were students from Stockholm School of Economics. They were approached through a friend of the author. The users were not previously known to the author nor did anyone have any previous knowledge of the Russian language.

4.1.3 The wizard of Oz
As previously mentioned, in order to keep the focus on the actual test and not be bothered by any technical limitations, the decision whether or not the user pronounced the word correctly was controlled by a Wizard. The Wizard was placed on a chair away from the user, with a keyboard placed behind cover so that the user could not see the keyboard was pressed. After working with the set amount of words during the course of this thesis, as well as receiving a basic introduction to the language by a native speaker, the author felt the requirements were met for him to be the Wizard in such a test. Having listened to the words several times, both the recordings and said by a native speaker, the author was able to make the judgment whether it was a correct pronunciation or not within the purpose of this test. Thus the author was the Wizard during all of the tests.
4.2 Data collection

At the end of each exercise, the computer would tell the user how many correct words they had during the test at the end. These results were collected by the author in order to see if any significant results could be obtained. In the second user study, the amount of time spent in the training phase was also collected. This was collected in order to see if there was a significant difference in the time spent with the different setups. The results from the scoring and time spent on the exercises will be presented in chapter 5. After the users were finished with the exercise they were given a questionnaire in the form of a Google form to fill in on a laptop.

There were a total of ten questions in the questionnaire given with the ability to answer using a 1-5 scale, giving a low score if they strongly disagree with the statement and a high score if they totally agree with the statement. The questions were answered once for each of the different exercises. After each question and the three ratings of it, there was a possibility to add a free comment. After ten questions with scaled answers, the users were given two questions with free text answers.
regarding each system, totaling six free text questions. Finally the user was asked if he or she had any previous knowledge of Furhat. The questions in the questionnaire can be found in Appendix C. In the second user study, question 9 was removed due to no actual correction being done by the teacher in this setup.
Chapter 5

Results

The following sections will present the results from the first user study. First the scores from the tests after each exercise will be presented and analyzed. After that results from the questionnaire will be presented. After the results of the first user study have been presented, the results from the second user study will be presented, along with the answers to that questionnaire.

5.1 Calculations from the first user study

In order to see the significance in the test scores from the exercise, we first have to choose a method to evaluate if the difference in the results is statistically significant. The first thing needed to be done is to realize whether or not the data is normally distributed. In this case, the result is a smaller sample size and it is initially not assumed to be normally distributed. In order to verify this, a Shapiro-Wilk test\(^9\) was done on the data to verify that the test results cannot be considered normally distributed. Therefore a method where the data is not normally distributed must be used. Since the test includes multiple comparisons, as the User 1 did the Tests A, B and C, a method which takes multiple comparisons into account must first be used. The numbers are on a continuous level in the test scores, with a possible score of 0-9. Using the Friedman test\(^10\) will allow us to first consider if there is a significant difference in the results from the test. If the Friedman test shows that there is a significant difference between the tests, post hoc testing with Wilcoxon signed-rank tests on the different combinations of related groups can be run. This however requires a Bonferroni correction\(^11\) because you are making multiple comparisons. The calculations before any corrections were done with an assumed alpha-value of 0.05. The actual test scores from the test can be found in Appendix D, this section will only include the results of the calculations.

5.1.1 Test scores

Evaluating the test scores in the Friedman test, the results were as follows where:

Test A is with a non-embodied voice.
Test B is with an animated face.
Test C is with Furhat.

---

\(^9\) https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test
\(^10\) https://en.wikipedia.org/wiki/Friedman_test
\(^11\) https://en.wikipedia.org/wiki/Bonferroni_correction
The higher the rank, the better the individual user performed in the test in relation to the other tests of the same user. Table 5.1 shows us that the mean rank is higher for the C across the user tests, and the calculations follow that there is a statistically significant difference in the test results depending on the setup, given by $\chi^2(2) = 13.43$ and $p = 0.0012$.

In order to verify where those differences lay, post hoc Wilcoxon signed-rank tests were performed on the different combinations of related groups:

Case 1. Non-embodied voice to animated face.
Case 2. Non-embodied voice to the robotic head Furhat.
Case 3. Animated face to the robotic head Furhat.

The Bonferroni correction must be applied, this giving a value of $0.05 / 3 = 0.017$ – meaning that a p-value larger than 0.017 does not indicate a statistically significant result.

Case 1 gives a p-value of 0.09102. The difference is not significant at $p \leq 0.017$.
Case 2 gives a p-value of 0.00222. The difference is significant at $p \leq 0.017$.
Case 3 gives a p-value of 0.01140. The difference is significant at $p \leq 0.017$.

Given the results, one can see that the difference in the results of using Furhat compared to the other was statistically significant. Comparing using a non-embodied voice to an animated face gave no statically significant difference. Given the mean rank from the Friedman test one can see that the users performed better in the exercise with Furhat compared to the other exercises. Looking at the mean scores from across all platforms, they indicate the result of higher performance using Furhat.
### Data summary

<table>
<thead>
<tr>
<th></th>
<th>Test</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samples</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>6.4</td>
<td>7</td>
<td>7.8</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>1.6857</td>
<td>1.1429</td>
<td>0.7429</td>
<td></td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>1.2984</td>
<td>1.069</td>
<td>0.8619</td>
<td></td>
</tr>
<tr>
<td>Std. Error</td>
<td>0.3352</td>
<td>0.276</td>
<td>0.2225</td>
<td></td>
</tr>
</tbody>
</table>

*Table 5.2: Data summary of the test results.*

### 5.1.2 Questionnaire answers

As mentioned the questionnaire answers were presented with the possibility on answering on a scale from 1 to 5, giving a low score if they strongly disagree with the statement and a high score if they totally agree with the statement. Since the answers are on an ordinal scale, using the mean values may not give a proper indication of the answers; therefore the median will be displayed instead. Since the questions are the same, but answered one time for each setup, they are compared again with the Friedman test to see if there is a statistical significant difference between the answers.

<table>
<thead>
<tr>
<th>Question</th>
<th>Samples</th>
<th>Median Case A</th>
<th>Median Case B</th>
<th>Median Case C</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>I found the system useful for learning new words.</td>
<td>15</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>I found the system entertaining.</td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>I found the system educational.</td>
<td>15</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>0.0009</td>
</tr>
<tr>
<td>I found the system more educational than practicing alone.</td>
<td>15</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>I found myself engaged in the system.</td>
<td>15</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>I could learn to understand new words through a system like this.</td>
<td>15</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.7047</td>
</tr>
<tr>
<td>I could learn how to pronounce new words through a system like this.</td>
<td>15</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>I felt encouraged to speak during the exercise.</td>
<td>15</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
It felt OK to be corrected by the computer during this exercise.

<table>
<thead>
<tr>
<th></th>
<th>15</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>0.0002</th>
</tr>
</thead>
</table>

Table 5.3: Results from the survey. The first column is the question. The second column is the amount of answers. The third, fourth and fifth column show the median of the answers. The final column show the calculated p-value of each question. The question where the differences between the answers were considered statistically insignificant is marked with a bold text. (When p > 0.05).

5.1.3 Free text comments

At the end of the questionnaire, two free text questions about each of the setups were given, totaling in six free text questions. Not all participants answered these, but continued on with them empty. The following sections will present an extract of these answers, the full list of answers can be found in Appendix F.

5.1.3.1 Positive feedback with the non-embodied voice

A few of the responses focused on what possibilities this could have for people suffering from any handicap. “It is good if you have a hard time reading on your own” and “It helps people who cannot read”, one comment even regarded the system positive if you find yourself having a hard time picking up new words from reading a book or something similar, “It is a way to learn if you have a hard time doing it by yourself from a book”. The comments about the positive feedback were generally centered on the fact that you could learn new words using such a setup.

5.1.3.2 Positive feedback with the animated face

The responses regarding the animated face seem to move away from the fact that this would be helpful for people with any kind of impairment. In general people seemed to express positive attitude toward the fact that the animated face could be of some help. Comments “It gives you some more personal touch than just a robot voice reading” and “Its an easy way to hear and learn new words, the face helped to focus” illustrate this, but there were also comments expressing an uncomfortable setting just speaking to a face on the screen, “The face was abit scary, i dont like talking to the screen” and “It was more personal than with no face, but you still feel pretty strange talking to the screen”.

5.1.3.3 Positive feedback with Furhat

The responses regarding Furhat seem to focus on the fact that moving the animated face onto an actual 3D-mask projection brings another dimension of life to the teacher in the exercise. Comments centered around “It felt natural to talk to a robot and not just a screen” and “Furhat gave the feeling
of a teacher sitting just with you, making you feel like you really have to learn and listen to him.”. This was not always expressed in a clearly positive way; “Furhat makes it feel like someone is watching you learn and you must do better” gives the impression that having Furhat created a more tense environment giving performance anxiety.

5.1.3.4 Negative feedback with the non-embodied voice
The negative feedback given around the setup with the non-embodied voice seems to center around the fact that it was not very entertaining. A lot of the comments mention that you either “lose focus after a while” and “It started to get pretty boring”. One comment even indicated that this setup is meant to be used by people suffering from impairment as the question was simply answered “cannot say, its not for me”.

5.1.3.5 Negative feedback with the animated face
The feedback on the animated face seems to revolve around the fact that even though a face is present on the screen, the personal connection is still missing. Comments like “The face looked scary at times, just staring at nothing” and “I felt no connection with the robot face” indicates this. Generally the help they got from having the face was not fully appreciated, with comments like “the face wasnt helping any more than with no face” and “the face was not good at helping with the word, he just showed it again”.

5.1.3.6 Negative feedback with Furhat
The comments are to some extent related the what was negative about the face, “He didnt help other than to tell me I was wrong and just repeat the word”, where you would hope and perhaps expect some more detailed feedback. Otherwise there seem to be a difference in the negative feedback, the negative feedback around Furhat has signs of people wanting to do more with Furhat. “I couldnt ask him my own questions”, “It couldnt answer when I had a question on my own” and “He couldnt talk any more than the other systems” goes to show that perhaps expectations from having a robotic face in front of you caused you to believe it would be a different interaction. People appear to perhaps be more open to interacting in a more humane way with the robotic head rather than just the face on screen.

5.2 Calculations from the second user study
The exercise in the second user study is structured in the same way as the first one, allowing the same principles of calculation to be applied to the data collected in the second user study. The actual
test scores from the test can be found in Appendix E, this section will only include the results of the calculations.

5.2.1 Test scores
Again, evaluating the test scores in the Friedman test, the results were as follows where:
Test A is with a non-embodied voice.
Test B is with an animated face.
Test C is with Furhat.

<table>
<thead>
<tr>
<th>Mean rank for tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>1.8</td>
</tr>
</tbody>
</table>

Table 5.4: Mean rank from Friedman test

Table 5.4 shows us that the mean rank is very similar across the user tests, and the calculations follows that there is no statistically significant difference in the test results depending on the setup, given by $x^2(2) = 0.65$ and $p = 0.7225$. Given that the Friedman test shows no statistically significant difference between the setups, no post hoc tests to identify the difference should be run. The test scores simply show that the different setups gave no significant difference in the test results.

<table>
<thead>
<tr>
<th>Data summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
</tr>
<tr>
<td>Samples</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Variance</td>
</tr>
<tr>
<td>Std. Deviation</td>
</tr>
<tr>
<td>Std. Error</td>
</tr>
</tbody>
</table>

Table 5.5: Data summary of the test results.

5.2.2 Time spent with the different setups
In the second user study, the user himself decided when he or she was ready to take the test. Thus it gave an opportunity to record the amount of time spent in the training phase with the different stations. Yet again the data is not normally distributed and a Friedman test will help identify if there is a significant difference in the time spent during the different exercises.
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.5</td>
<td>1.6</td>
<td>2.9</td>
</tr>
</tbody>
</table>

*Table 5.6: Mean rank from Friedman test*

The higher the rank, the longer the individual user spent in the test situation in relation to the other tests with the same user. Table 5.5 shows us that the mean rank is higher for the C across the user tests, and the calculations follows that there is a statistically significant difference in the test results depending on the setup, given by $x^2(2) = 13.82$ and $p = 0.001$.

In order to verify where those differences lay, post hoc Wilcoxon signed-rank tests were performed on the different combinations of related groups:

Case 1. Non-embodied voice to animated face.
Case 2. Non-embodied voice to the robotic head Furhat.
Case 3. Animated face to the robotic head Furhat.

The Bonferroni correction must be applied, this giving a value of $0.05 / 3 = 0.017$ – meaning that a $p$ value larger than 0.017 does not indicate a statistically significant result.

Case 1 gives a $p$-value of 0.13888. The difference is not significant at $p \leq 0.017$.
Case 2 gives a $p$-value of 0.00512. The difference is significant at $p \leq 0.017$.
Case 3 gives a $p$-value of 0.00512. The difference is significant at $p \leq 0.017$.

Given the results, one can see that the difference in the time spent using Furhat compared to the other was statistically significant. Comparing using a non-embodied voice to an animated face gave no statically significant difference. Given the mean rank from the Friedman test one can see that the users spent a longer time in the exercise with Furhat compared to the other exercises. Looking at the mean scores from across all platforms, they indicate the same results.
### 5.2.3 Questionnaire answers in second study

The questionnaire from the second study was the same as in the first study, apart from having two less questions. Therefore, the results of the questionnaire will be presented in the same way. Since the questions are the same, but answered one time for each setup, they are compared again with the Friedman test to see if there is a statistical significant difference between the answers.

<table>
<thead>
<tr>
<th>Question</th>
<th>Samples</th>
<th>Median Case A</th>
<th>Median Case B</th>
<th>Median Case C</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>I found the system useful for learning new words.</td>
<td>11</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.0699</td>
</tr>
<tr>
<td>I found the system entertaining.</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>I found the system educational.</td>
<td>11</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>0.0104</td>
</tr>
<tr>
<td>I found the system more educational than practicing alone.</td>
<td>11</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>I found myself engaged in the system.</td>
<td>11</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>I could learn to understand new words through a system like this.</td>
<td>11</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.6219</td>
</tr>
<tr>
<td>I could learn how to pronounce new words through a system like this.</td>
<td>11</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>I felt encouraged to speak during the exercise.</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0.0801</td>
</tr>
</tbody>
</table>

*Table 5.8: Results from the survey. The first column is the question. The second column is the amount of answers. The third, fourth and fifth column shown the median of the answers. The final column*
show the calculated p-value of each question. The question where the differences between the answers were considered statistically insignificant is marked with a bold text. (When p> 0.05).

5.2.4 Free text answers
The same questionnaire was used for the second study, which included free text questions at the end. The following sections excerpts will be presented from the free text answers, the full answers can be found in Appendix G.

5.2.4.1 Positive feedback with the non-embodied voice
The feedback given in this question was a lot about the fact that it’s a flexible and different way of learning new words, rather than just reading a book. Comments like “Great if you don’t enjoy learning by reading a word list” and “This could be more fun than just reading a book” expresses that. The flexibility in the training was highlighted in comments such as “It’s a self-controlled lesson, you set the pace” and “It’s an adaptive training session, everyone can use it”. In general the positive feedback seemed to center around the fact that it could be more fun if you do not enjoy learning by reading a word list and the fact that it is adaptive to the user.

5.2.4.2 Positive feedback with the animated face
The feedback given to the animated face was similar to the same feedback given to the exercise with the non-embodied voice. But the added experience of the face seemed to be regarded as a positive effect on learning the pronunciation of the word. “Helps you not only hear the word, but you also get to see it” and “Gives a better introduction to the pronunciation”. Also comments about the face making the experience more fun and living experience were given, “Even more fun with a face, almost like a teacher” and “The face gives a more living experience”.

5.2.4.3 Positive feedback with Furhat
Finally the feedback given to Furhat continued in a positive fashion, with Furhat perhaps given a slightly better experience with regards to the same topics. “Gives the best introduction to pronunciation” and “The robot helps you with how the word should be said, just as the face, but its more fun” points toward the same positive experience from the previous system, but Furhat is regarded as more fun than just the face. “The robot gives you an actual teacher to student feeling”
and “The robot feels like a nice and friendly teacher” gives an indication that perhaps a different personal connection is established with Furhat.

5.2.4.4 Negative feedback with the non-embodied voice
The negative feedback with the non-embodied voice seems to center on the fact that it could perhaps be pretty inefficient and slow if you would want to learn a lot. “It might feel slow for you”, “It's pretty inefficient if you want to learn many words” and “It would be boring after a while” points toward this.

5.2.4.5 Negative feedback with the animated face
The negative feedback with the animated face was very much the same as with the non-embodied voice, many of the answers even referred to the answers to the previous question, or simply expressed the same thing as the previous answer. “Same as before”, “It is also inefficient” and “This too could be boring” shows this. The feedback is shared on this setup as with the one before – it is perhaps not a very efficient way of learning new words, and it could get boring after a while.

5.2.4.6 Negative feedback with Furhat
Again, the negative feedback was very much the same as with the previous setups. Answers were again referring to the answer to previous questions, or simply repeating the previous answer – “This too might feel slow for you”, “Same as before” and “It is also inefficient” all show this. Again the feedback is shared on this setup as the ones before.
Chapter 6

Discussion

This chapter discusses the outcome of the results in the user studies in an attempt to see any meaning to these results.

6.1 Interpreting results

In the first study, looking at the test scores and the statistical analysis of them we can actually see that there is a statistically significant difference between the different exercises. Trying to pinpoint them further with Wilcoxon signed-rank method with adjusted p-values, we see that the difference can actually be found when you take the step from the screen to the actual 3D world. Even if the same facial animations are displayed on an animated face on your screen, moving out to a robotic face yields better individual scores in relation to the other exercises. This however does not hold in the second user study, where the differences in results cannot be considered statistically significant. It appears the maximum time given in the training part was more than enough for the user to learn more or less all the words, as the results were high across all setups and there was plenty of time left. What could be observed though is that more time was spent with Furhat than with the other setups.

This can perhaps be related to what can be seen in theories around motivation. The motivation for doing a task has the simple additive factor of task motivation, where the task performed is enjoyable or fun. Because of the fact that the experience of playing along with an actual robot face is a more technologically advanced feature, it can likely be assumed that such an experience has a higher news value to the user. Because the experience is new and perhaps unique, this may affect the user in a way which could increase the interest in the task and thus increase the task motivation, giving a better learning experience. In the second user study, the results were unaffected and this is likely due to the fact that the maximum time given was clearly enough across all systems to learn all the words. But in the first user study, where the user was introduced to the word only one time, the results were better with Furhat. Perhaps the task motivation of doing the exercise with Furhat was higher and thus increasing the concentration and focus when you are being introduced to the words, allowing you to remember them better. Just as it can be observed that more time was spent training with Furhat than with the other system, simply perhaps because they enjoy it more. In other words, having a higher task motivation, simply enjoying something more, could affect the learning process in
a positive way. In this study it could be the explanation to the higher results achieved with Furhat in the first user study, and longer time spent with Furhat in the second study.

The role of the computer in the two different exercises is notably different. In the first exercise, the computer takes an active role as the teacher going through the exercise together with the user. In the second exercise, the computer acts more as an introducer on command and an examiner. During the training in the second exercise the user is more active and one could question the need for the interaction with the agent. The results, however, perhaps show that even though the teacher role is more passive the effect of having Furhat still gives a higher task motivation.

Some of the users expressed, through the questionnaire and in passing after the exercise was finished, that they felt a stronger personal connection with Furhat and was hoping for more dialogue to be exchanged with him. This goes to show that perhaps it is not only the news value and fun in the new experience with a robotic face, but perhaps the robotic face moves towards a more human touch causing you to be more emotionally invested in the experience. The situation could be similar to sitting face-to-face with a teacher and being taught and expected to learn. The extrinsic motivation could also come in to play in such a situation, where you simply are more afraid of the punishment of a bad grade, in this case a bad score on the test coming from something you have a much more human like connection with.

6.2 Questionnaire

The answers to the questions all point towards a higher perceived experience with Furhat in many of the aspects in the questions – fun, educational, engaging. In the first study, all questions but one we could see a statistically significant difference in the ratings, and the higher ratings were given to the experience with the robotic head. The only question where the answers were too similar was about whether or not the system could help you understand new words through such an exercise. In the sense that all of them introduced the translation of 9 new words, it is perhaps reasonable that the feeling of learning new words were similar across all exercises. But when actual pronunciation was a factor, the exercise with no face was rated lower in regards to the ability of such an exercise helping out with pronunciation.

In the second study, the majority of the questions also points toward a higher perceived experience in the previously mentioned aspects. The experience of the different setups being useful for learning new words was the same across all systems. The second study actually left the control of the training
to the user, which could explain the feeling of the training potential being the same across all systems – at the end of the day you are in control yourself. The same result with learning new words persisted. The final question regarding whether or not the user felt encouraged to speak was considerably lower and same across all setups in the second user study. This is most likely explained due to the fact that the actual exercise was not a flowing dialogue in the same way as the first one, this was controlled more by the user and the computer answered when user prompted for translations of words. If there is less dialogue in general, there is less encouragement to speak.

As explained in the previous section, the computer had a much passive role in the second exercise. The answers from the questionnaire still points toward a higher enjoyment with Furhat, even though the exercise is more user driven.

The question regarding if it felt okey to be corrected by the system showed a higher acceptance towards being correct by Furhat. It even feels better to be corrected by an animated face than by a non-embodied voice. Perhaps this points towards the fact that having some sort of connection to whomever you are doing the exercise with affects how you feel about the interaction. In general though, the exercise involving Furhat seemed to receive higher ratings in all of the aspects. Whether this is affected by the fact that it could be much more interesting because it is newer technology can only be speculated on in this case. The fact remains though, the perceived experience of more fun, motivation and engagement resulted in a higher score in one study and more time spent in the exercise in the other.
Chapter 7
In this chapter the final conclusions regarding the work in this thesis will be presented along with thoughts about future work within this field.

7.1 Conclusion
The scientific question the author wanted to explore was:

How is the language learning process of vocabulary learning affected by the physical presence of a robotic head with a 3D face?

After going through a literature study on different applications of computers in language learning, it was decided upon a design of an exercise which could point towards some sort of answer to this question. Through the use of an implemented Russian vocabulary exercise ready to run with three different setups, the user study with 15 participants was held. The user study could show a higher score on the test at the end of the exercise when going through the exercise with Furhat. In order to get more data to analyze and discuss, the exercise was reformed to focus on the active vocabulary rather than the passive. The results from the second user study with 11 users showed no significant difference in actual test scores from the exercises, but it showed that users spent more time training with Furhat than the other systems. This could indicate a higher task motivation, simply because the users think it is more fun to exercise with Furhat, causing them to train longer with Furhat before they get bored. Increased task motivation could also explain the higher results from the first user study, as higher task motivation could affect the learning in a positive way. Since the users were shown the words only one time each in the first user study, the increased motivation could affect your ability to recall the word after just hearing it once.

The questionnaire following the exercises would also point towards a greater enjoyment and the feeling of a better educational environment when working with Furhat, a physically present robotic head.

7.2 Future work
After speaking briefly to the test participants and looking at some of the answers on the questionnaire, it shows that people are interested in having a more extensive dialogue with the robot. With the animated face, people did not feel the same personal connection leaving them uninterested in having further exchange in dialogue. Perhaps future work could include having a more open dialogue during the exercise, and the material taught not only bound to vocabulary, but
full grammatical sentences. The expressed feeling of being more interested in speaking with Furhat than the other systems could perhaps show a better scoring in an exercise not only bound to one word at the time.

In general, perhaps looking at what physical and biological effects the presence of the robot has would contribute to understanding the results. Monitoring the heart rate, the amount of sweating and such methods could be used to monitor for any difference. The aspect of time could be expanded further than done in the second study – perhaps one could see if the presence of the robot allows for the exercises to be completed faster. Having a better time limit than what was given in the second user study could perhaps show a difference too as the time given in this study was enough to achieve high scores across all setups.

Working with virtual teachers in this thesis has given an insight to the strengths and benefits computers can bring in educational aspects. Even though this thesis was centered only on the use of computers in an educational environment, not involving any comparisons with exercise with a human teacher rather than a virtual, the results are in many ways positive. The users in this study had different backgrounds and different experience with computers, but they all showed the ability of quickly and effectively understanding how to work with computers in subjects where they have perhaps not done so before. The author believes this is positive in the larger sense that introducing virtual teachers as a resource where the human resources simply are not enough could greatly benefit society from an educational perspective. If not in the form of a robotic head, but at least a virtual teacher in the form of software could complement existing resources in order to continue to provide an educational environment for all individuals in school.

There is much to see as the introduction of physically present robots approaches our everyday life. This study is scratching the surface of what is to come, and while currently the news value of a robotic head could be enough to achieve a higher task motivation, this is perhaps not as relevant in the future when robots could become a natural part of society.
References


Appendices
Appendix A
List of words used
English word – Russian translation in Latin alphabet.

Set 1
Black – chernyy
Blue – siniy
Brown – korichnevyy
Green – zelenyy
Pink – rozovyy
Purple – fioletovyy
Red – krasnyy
White – belyy
Yellow – zheltyy

Set 2
Carpet – kover
Bathtub – vanna
Door – dver’
Key – klyuch
Armchair – kreslo
Stairs – lestnitsa
Roof – krysha
Shoes – obuv’
Window – okno

Set 3
Cat – kot
Chicken – kuritsa
Cow – korova
Dog – sobaka
Elephant – slon
Fish – ryba
Horse – loshad’
Penguin – pingvin
Pig – svin’ya
Appendix B

Definition of gestures

<gesture name="smile">
    <param name="SMILE_CLOSED">16(0.5) 16(0.5) 16(0)</param>
    <param name="BROW_UP_LEFT">16(1) 16(1) 16(0)</param>
    <param name="BROW_UP_RIGHT">16(1) 16(1) 16(0)</param>
</gesture>

<gesture name="smilewink">
    <param name="SMILE_OPEN">16(1) 16(0)</param>
    <param name="BLINK_LEFT">16(1) 16(0)</param>
</gesture>

<gesture name="smileopen">
    <param name="SMILE_OPEN">16(1) 16(1) 16(1) 16(0)</param>
</gesture>
Appendix C

Questions in the questionnaire
Q1. I found the system useful for learning new words.
1-5 with a non-embodied voice
1-5 with an animated face
1-5 with Furhat.

Q2. I found the system entertaining.
1-5 with a non-embodied voice
1-5 with an animated face
1-5 with Furhat.

Q3. I found the system educational
1-5 with a non-embodied voice
1-5 with an animated face
1-5 with Furhat.

Q4. I found the system more educational than practicing alone
1-5 with a non-embodied voice
1-5 with an animated face
1-5 with Furhat.

Q5. I found myself engaged in the system.
1-5 with a non-embodied voice
1-5 with an animated face
1-5 with Furhat.

Q6. I could learn to understand new words through a system like this.
1-5 with a non-embodied voice
1-5 with an animated face
1-5 with Furhat.

Q7. I could learn how to pronounce new words through a system like this.
1-5 with a non-embodied voice
1-5 with an animated face
1-5 with Furhat.

Q8. I felt encouraged to speak during the exercise.
1-5 with a non-embodied voice
1-5 with an animated face
1-5 with Furhat.
Q9. It felt OK to be corrected by the computer during this exercise.
1-5 with non-embodied voice
1-5 with an animated face
1-5 with Furhat.

Q10 What was positive about the system...
..with a non-embodied voice? -- free text answer
..with animated face? – free text answer
..with Furhat? – free text answer

Q11 What was negative about the system...
..with a non-embodied voice? -- free text answer
..with animated face? – free text answer
..with Furhat? – free text answer
# Appendix D

## Test scores

<table>
<thead>
<tr>
<th>Inget (animals)</th>
<th>ansikte (colours)</th>
<th>furhat (house)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Furhat (colours)</td>
<td>inget (house)</td>
<td>ansikte (animals)</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ansikte (house)</td>
<td>inget (colours)</td>
<td>furhat (animals)</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inget (colours)</td>
<td>ansikte (house)</td>
<td>furhat (animals)</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Furhat (house)</td>
<td>ansikte (colours)</td>
<td>inget (animals)</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ansikte (animals)</td>
<td>furhat (colours)</td>
<td>inget (house)</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inget (house)</td>
<td>furhat (animals)</td>
<td>ansikte (colours)</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Furhat (colours)</td>
<td>inget (animals)</td>
<td>ansikte (house)</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ansikte (animals)</td>
<td>Furhat (house)</td>
<td>inget (colours)</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inget (animals)</td>
<td>furhat (colours)</td>
<td>ansikte (house)</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Furhat (house)</td>
<td>ansikte (animals)</td>
<td>inget (colours)</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>ansikte(colours)</td>
<td>inget(house)</td>
</tr>
<tr>
<td>----</td>
<td>-----------------</td>
<td>-------------</td>
</tr>
<tr>
<td>6</td>
<td>ansikte(colours)</td>
<td>inget(house)</td>
</tr>
<tr>
<td>13</td>
<td>inget(colours)</td>
<td>Furhat(animals)</td>
</tr>
<tr>
<td>14</td>
<td>furhat(house)</td>
<td>ansikte(colours)</td>
</tr>
<tr>
<td>15</td>
<td>ansikte(animals)</td>
<td>inget(colours)</td>
</tr>
<tr>
<td>7</td>
<td>ansikte(animals)</td>
<td>inget(colours)</td>
</tr>
</tbody>
</table>
## Appendix E

<table>
<thead>
<tr>
<th></th>
<th>Inget (animals)</th>
<th>ansikte (colours)</th>
<th>furhat (house)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9, 191</td>
<td>7, 216</td>
<td>9, 257</td>
</tr>
<tr>
<td>2</td>
<td>Furhat (colours)</td>
<td>8, 232</td>
<td>ansikte (animals) 8, 214</td>
</tr>
<tr>
<td>3</td>
<td>ansikte (house)</td>
<td>9, 234</td>
<td>furhat (animals) 9, 270</td>
</tr>
<tr>
<td>4</td>
<td>Inget (colours)</td>
<td>ansikte (house)</td>
<td>furhat (animals) 9, 227</td>
</tr>
<tr>
<td>5</td>
<td>Furhat (house)</td>
<td>ansikte (colours)</td>
<td>Inget (animals) 9, 214</td>
</tr>
<tr>
<td>6</td>
<td>ansikte (animals)</td>
<td>furhat (colours)</td>
<td>Inget (house) 9, 190</td>
</tr>
<tr>
<td>7</td>
<td>Inget (house)</td>
<td>furhat (animals)</td>
<td>ansikte (colours) 9, 300</td>
</tr>
<tr>
<td>8</td>
<td>Furhat (colours)</td>
<td>ansikte (animals)</td>
<td>Inget (house) 8, 204</td>
</tr>
<tr>
<td>9</td>
<td>ansikte (animals)</td>
<td>9, 240</td>
<td>Inget (colours) 8, 250</td>
</tr>
<tr>
<td>10</td>
<td>Inget (animals)</td>
<td>Furhat (house)</td>
<td>ansikte (house) 7, 178</td>
</tr>
<tr>
<td>11</td>
<td>furhat (house)</td>
<td>ansikte (animals)</td>
<td>Inget (colours) 7, 224</td>
</tr>
</tbody>
</table>
### Positive feedback with non embodied voice

<table>
<thead>
<tr>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>It can help people learn new words and how to say them</td>
</tr>
<tr>
<td>I did not enjoy it</td>
</tr>
<tr>
<td>It could help me learn some new words, but it was pretty boring</td>
</tr>
<tr>
<td>It is good if you have a hard time reading on your own</td>
</tr>
<tr>
<td>Its an easy way to hear and learn new words</td>
</tr>
<tr>
<td>nothing special</td>
</tr>
<tr>
<td>It is a way to learn if you have a hard time doing it by yourself from a book</td>
</tr>
<tr>
<td>It is one way of learning new words</td>
</tr>
<tr>
<td>It can help you learn if you are not motivated</td>
</tr>
<tr>
<td>I helps people who cannot read</td>
</tr>
</tbody>
</table>

### Positive feedback with an animated face

<table>
<thead>
<tr>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>It can help you learn new words and more dialog when you have someone to talk to</td>
</tr>
<tr>
<td>it was more personal than with no face, but you still feel pretty strange talking to the screen</td>
</tr>
<tr>
<td>It gave something to look at apart from the pictures and word</td>
</tr>
<tr>
<td>It gives you some more personal touch than just a robot voice reading</td>
</tr>
<tr>
<td>its an easy way to hear and learn new words, the face helped to focus</td>
</tr>
<tr>
<td>you felt a bigger connection doing it with a teaching face</td>
</tr>
<tr>
<td>The face can help you be more concentrated</td>
</tr>
<tr>
<td>The face helped keep the feeling that you were being taught something, and therefor made you concentrate</td>
</tr>
<tr>
<td>It can also help you if you are not motivated</td>
</tr>
<tr>
<td>The face was abit scary, i dont like talking to the screen.</td>
</tr>
</tbody>
</table>

### Positive feedback with an animated face

<table>
<thead>
<tr>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Furhat makes it feel like someone is watching you learn and you must do better</td>
</tr>
<tr>
<td>It felt natural to talk to a robot and not just a screen</td>
</tr>
<tr>
<td>It felt more alive to talk to, i would want to have longer conversations with such a robot, not just answer with one word</td>
</tr>
<tr>
<td>Gives a personal touch and a want to cooperate</td>
</tr>
<tr>
<td>Its an easy way to hear and learn new words in a more fun way</td>
</tr>
<tr>
<td>It was easier to concentrate when you had to talk to the robot</td>
</tr>
<tr>
<td>The robot makes you focus alot more and gives nicer feedback</td>
</tr>
<tr>
<td>Furhat gave the feeling of a teacher sitting just with you, making you feel like you really have to learn and listen to him.</td>
</tr>
<tr>
<td>It can also help you if you are not motivated, the robot feels like a personal teacher</td>
</tr>
</tbody>
</table>
### Negative feedback with non embodied voice

- The feedback was unpersonal
- It got a bit slow and boring
- It started to get pretty boring
- It was too boring
- It feels boring looking at a screen after a while
- It felt uninspiring
- Cannot say, it's not for me
- I could have learned it faster myself
- You lose focus after a while

### Negative feedback with an animated face

- Same
- The face did not feel personal or helping
- It was pretty slow
- I felt no connection with the robot face
- The face looked scary at times, just staring at nothing
- The face was not good at helping with the word, he just showed it again
- The face wasn't helping any more than with no face
- It didn't help too much with how it was supposed to be pronounced
- The face didn't feel as personal as the robot

### Negative feedback with Furhat

- Same
- I couldn't ask him my own questions
- Lack of things to say to the robot
- Better targeted feedback to me, so I know how to improve
- He didn't help other than to tell me I was wrong and just repeat the word
- He couldn't talk any more than the other systems
- At times it felt scary when he was staring at you
- He could have given more clear feedback instead of just saying I was wrong and showing me how to say the word again
- It couldn't answer when I had a question on my own
### Appendix G

#### Positive feedback with non embodied voice

<table>
<thead>
<tr>
<th>Feedback</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Its a self-controlled language lesson, you set the pace</td>
<td>You decided yourself when you were ready with the practice</td>
</tr>
<tr>
<td>Helps you at least hear the words you’re training, compared to when you</td>
<td>Its good cause it focuses on learning and speaking the words instead</td>
</tr>
<tr>
<td>only read them yourself</td>
<td>of just learning them in writing</td>
</tr>
<tr>
<td>Great if you don’t enjoy learning by reading a word list</td>
<td>Feels like it could be beneficial for children, like an interactive</td>
</tr>
<tr>
<td>This could be more fun than just reading a book</td>
<td>“point book”</td>
</tr>
<tr>
<td>Its an adaptive training session, everyone can use it</td>
<td></td>
</tr>
</tbody>
</table>

#### Positive feedback with an animated face

<table>
<thead>
<tr>
<th>Feedback</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>The same as the answer before, but you also had a face to see the words</td>
<td>Helps you not only hear the word, but you also get to see it</td>
</tr>
<tr>
<td>being said</td>
<td>Gives a better introduction to the pronunciation</td>
</tr>
<tr>
<td>The face gives a more living experience</td>
<td>Having the face on screen helps to see how the word should be said</td>
</tr>
<tr>
<td>Even more fun with a face, almost like a teacher</td>
<td>You learn more when it comes to saying the words</td>
</tr>
</tbody>
</table>

#### Positive feedback with an animated face

<table>
<thead>
<tr>
<th>Feedback</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>The most optimal way, the robot face helps a lot with how the words</td>
<td>See previous answer</td>
</tr>
<tr>
<td>should be said</td>
<td>The best way of seeing and hearing words, the face gives the best</td>
</tr>
<tr>
<td></td>
<td>connection</td>
</tr>
<tr>
<td></td>
<td>Gives the best introduction to pronunciation</td>
</tr>
<tr>
<td></td>
<td>The robot gives you an actual teacher to student feeling</td>
</tr>
<tr>
<td></td>
<td>The robot helps with how the word should be said, just as the face,</td>
</tr>
<tr>
<td></td>
<td>but its more fun</td>
</tr>
<tr>
<td></td>
<td>The robot feels like a nice and friendly teacher</td>
</tr>
<tr>
<td></td>
<td>You learn more when it comes to saying the words</td>
</tr>
</tbody>
</table>

#### Negative feedback with non embodied voice

<table>
<thead>
<tr>
<th>Feedback</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>It might feel too slow for you</td>
<td>It gets pretty repetitive</td>
</tr>
<tr>
<td></td>
<td>Its pretty inefficient if you want to learn many words</td>
</tr>
<tr>
<td></td>
<td>Not so much focus on the spelling</td>
</tr>
<tr>
<td></td>
<td>It would be boring after a while</td>
</tr>
<tr>
<td></td>
<td>Feels like it would be better for children than for adults</td>
</tr>
<tr>
<td></td>
<td>If you learn better from a book, you don’t need this</td>
</tr>
</tbody>
</table>

60
I wanted a way list of words at the same time, the only way I could see the spelling was to click on the picture. I want a list with all the words at all times.

**Negative feedback with an animated face**

- It might feel too slow for you
- Same as before
- It is also inefficient
- Not much spelling either
- This too could be boring

Not much, the face adds the pronunciation help
Same as previous question, I was missing a word list.

**Negative feedback with Furhat**

- This too might feel slow for you
- Same as before
- It is also inefficient
- No spelling either
- This too could be boring

No special negative feedback
Same as last answer.