Multimodal Technologies and Interaction

MDPI

Article

Robot-Enhanced Language Learning for Children in Norwegian Day-Care Centers

Till Halbach, Trenton Schulz, Wolfgang Leister * and Ivar Solheim

Norsk Regnesentral (Norwegian Computing Centre), 0373 Oslo, Norway; Till.Halbach@nr.no (T.H.); Trenton.Schulz@nr.no (T.S.); vrslhm@gmail.com (I.S.)
* Correspondence: Wolfgang.Leister@nr.no

Abstract: In a case study, we transformed the existing learning program Language Shower, which is used in some Norwegian day-care centers in the Grorud district of Oslo municipality, into a digital solution using an app for smartphones or tablets with the option for further enhancement of the presentation by a NAO robot. The solution was tested in several iterations and multiple day-care centers over several weeks. Measurements of the children's progress across learning sessions indicated a positive impact of the program using a robot as compared to the program without a robot. In situ observations and interviews with day-care center staff confirmed the solution's many advantages, but also revealed some important areas for improvement. In particular, the speech recognition needs to be more flexible and robust, and special measures have to be in place to handle children speaking simultaneously.

Keywords: social robot; mobile app; e-learning; language education; kindergarten; preschool; case study

Citation: Halbach, T.; Schulz, T.; Leister, W.; Solheim, I. Robot-Enhanced Language Learning for Children in Norwegian Day-Care Centers. *Multimodal Technol. Interact.* **2021**, *5*, 74. https://doi.org/10.3390/mti5120074

Academic Editors: Sofia Serholt and Shruti Chandra

Received: 31 August 2021
Accepted: 9 November 2021
Published: 24 November 2021

1. Introduction

In the Grorud district of the city of Oslo, up to 80% of the children in the day-care centers come from a non-Norwegian speaking or immigrant background. These children often speak a different language at home. This introduces an issue when the children begin school as they are behind their peers when it comes to Norwegian vocabulary. In spite of measures to help improve children's language learning in the Grorud district's preschools, about 40% of the children starting school in 2017 needed additional Norwegian language support [1]. New methods were needed.

One area to improve was building Norwegian vocabulary. The pedagogic staff in the Grorud district of Oslo developed the "Language Shower" specifically to build the children's vocabulary [1]. The idea of the program is to "shower" the children with Norwegian words every day over a longer period, with the goal to strengthen their oral and vocabulary skills. Currently, the program lasts one year, primarily targeting children aged four to six. The program also has a secondary target of the children's parents, who in many cases are in need of language training as well. Overall, Language Shower is designed as a supplement to other language-learning measures in the district's day-care centers, and it is meant to be used in a group setting of six to eight children.

The municipality was interested in the possibility of using a social robot to play the role of the session leader in Language Shower sessions to assist the staff and increase the children's engagement. Social robots have recently become popular in a variety of applications ranging from entertainment to customer expedition to teaching and education [2]. We created a prototype for a robot-led Language Shower to examine how social robots can enhance and support children's learning of a second language in three day-care centers in the Grorud district. The development and trials were carried out as a case study between autumn 2018 and spring 2021, and the overall objective of the work was to enhance the Language Shower learning program in the district with a social robot. Our goal in this

paper is to report how the robot-led Language Shower compared to the original Language Shower in our case study and document the difficulties and experiences in implementing the robot-led Language Shower.

Parts of the study reported in this paper were based on a pilot study by Fuglerud and Solheim [3]. They used a NAO robot [4] in two studies with children with autism spectrum disorder and with children that were second language learners, observing increased participation and involvement in language learning. Fuglerud and Solheim [3] recommended investigating the role of the robot in the overall pedagogical approach and its ethical implications in further studies. Further, the themes of engagement and participation, the combination of tools, the learning outcomes, the need for better speech recognition, language issues, the session length and group size, personalization and an easy to use interface, and technical support are revisited in the current article.

We documented the initial work around this study earlier [5]; here, we go into detail around implementing the Digital Language Shower, the issues we faced, the actual study, and the results. The remaining article is straightforward. After the outlining background and related work (Section 2), we describe the components and content for the Digital Language Shower (Section 3). Then, we describe how the trials were conducted (Section 4) and discuss the trials' results (Section 6). Finally, we conclude the article with possibilities for future improvement (Section 8).

2. Background

A social robot has been defined as an "... autonomous or semi-autonomous robot that interacts and communicates with humans by following the behavioral norms expected by the people with whom the robot is intended to interact" ([6], p. 592). Social robots can be used in a variety of ways in an educational setting for both adults [7] and school children [8]. Examples for children as the target learning group include using social robots as tools for children's group work to increase participation and proactiveness [9,10],or enabling remote attendance, tele-teaching, where students can use an avatar to remotely attend classes and be present in the classroom [11,12]. Another application of social robots can be as tutors for stimulating scientific curiosity [13] or improving a child's engagement, motivation, and performance in subjects such as learning Sign Language [14]. We examine how social robots have been used for language education and how our work relates to the previous work. Then, we examine some of the issues that one needs to consider when designing a solution that uses a social robot. Afterwards, we discuss the Language Shower concept.

2.1. Social Robots in Language Instruction

An early example of a social robot in language instruction was a humanoid robot deployed in an elementary school in Japan where children could interact with a robot during a break after lunch [15,16]. To help reduce the complexity of the interactions, the children were told that the robot could only speak English. This had the effect that it motivated the children to learn and use English, which led to the children learning new English vocabulary.

In many of the language-learning projects, social robots have been used for learning English, German, Dutch, Italian, Spanish, Japanese, Korean, and Persian [17,18]. There are not many studies addressing the use of social robots for learning Norwegian, which is a language used by fewer people, other than the work we build upon [3,5].

Different strategies have been applied for different studies. For example, one study examined how a robot (a NAO) could help German children aged 4–7 *scaffold* their second language learning of English [19]. That is, a robot would provide temporary assistance (a scaffold) through explanations to train the children in English. If the children could successfully internalize the knowledge, the scaffold could be removed or adapted to the next area of training. The robot would also detect flagging motivation (via Wizard of Oz [20], that is a person observing through NAO), and attempt to re-engage the child. Not all children completed the training, but the study showed that the robot could often

re-engage children and that the explanations were helpful to the group of children that completed the training. This study showed that robots can be useful for re-engaging training with children to maintain motivation and help in learning the language.

A recent survey [21] examined a number of studies that were performed in Robot-Assisted Language Learning (RALL). The survey found that robots were often autonomous, teleoperated, or a combination of both. The robots served many different functions such as: being a conversation partner, an aid in building vocabulary, grammar, pronunciation, reading comprehension, writing clarity, or assessing language ability. Often, the robot helped with vocabulary building (especially for a second language) or as a conversation partner. The robots had different forms: human-like, animal-like, or machine-like, and they often took the role as an assistant to the teacher or a peer or tutor. Often, the learners in the studies had increased motivation and interest in language. However, the survey found limitations: the robots were normally limited to a few sessions; the context was mostly children in the classroom in Asia; the statistical power of the studies was low. Our Digital Language Shower follows a similar setup, building vocabulary as an assistant, and it adds an additional context of Norwegian day-care centers (a younger group than what was examined in the survey).

Another review [22] examined social robots for language learning, including studies focusing on word learning in preschool and young school-aged children. The review included 33 studies, 13 of them focusing on word learning, and considered them related to learning outcome. Regarding learning outcome, the review concluded that the learning gains were similar for social robots and human teachers. The learning gain seemed to be generally low in most of the studies since, as the survey [21] above, the research design of many studies only included one session. The review also considered effects, such as effects on the learners' motivation, the novelty effect, and social behavior. There appeared to be positive effects of social robots on learning-related emotions, but the novelty effect was often not considered in the reviewed studies, while other studies that did run longer reported a declining interest as time progressed.

From the review and survey, it appears that at least some results on learning-related emotions and learning gains might stem from the initial excitement (or novelty) of using the robot. Regarding social behavior, the existing evidence with respect to social robots seems to be mixed, as it could increase children's engagement in learning tasks, as well as distract them from learning. Still, there are a variety of issues to consider when using a social robot for language learning.

2.2. Issues and Challenges in Using Social Robots for Language Learning

One issue to consider when using social robots is the role the robot should have in the instruction. Determining and assigning roles in Human–Robot Interaction (HRI) has often been a complex and dynamic process [23]. One study reported that children seemed to learn better when a NAO robot played the role of a peer instead of a teacher [24]. In other studies, robots acted as teaching assistants in the classroom. For example, RoboSapiens assisted teachers for five weeks in an elementary school when they read stories using different voices, led recitals, provided friendly feedback, gave quizzes, and moved when students asked [25]. Similarly, a NAO robot was used with some children to read out vocabulary words with pictures shown on slides behind it and pantomimed the meanings, while also providing entertainment such as singing and dancing [8]. The results of the study showed that students learned faster and learned more words, compared to a control group.

Once a role is settled, another issue can be how to structure the interaction. The structure can consist of several elements. Should a session be a group or a one-on-one interaction? One-on-one interactions provide better opportunities for tailoring the experience to the child, but there are also benefits to learning in a group environment. One also needs to determine the optimal length of sessions, the number of words that should be taught per session, and the optimal age group for using robots. It is also important to consider how learning outcomes can be measured before, during, and after the robot session.

Another issue to consider is how the robot should behave socially during the instruction. A robot that behaves more seriously than playfully may fit better for serious tasks [26]. Although social behaviors can be effective at engaging students, they can distract children [27]. Furthermore, positive effects have been observed when personalizing the behavior of two autonomous robots that acted as learning companions for children [28].

One challenge mentioned often in studies and reviews was Automatic Speech Recognition (ASR) for children [3,17,18,21]. Word recognition of children's speech in current solutions was two- to five-times worse than adults' speech [21]. Others have examined issues surrounding child speech recognition in the human–robot interaction context [29]. The study examined how microphones and noise from the robot (an earlier version of NAO) and the surrounding area affected the recognition. Overall, better microphones raised the rate of recognition significantly, but it required that the audio be normalized before recognition, something that happened in the robot's built-in microphone automatically. The additional noise of the robot's servos did not seem to affect the recognition rate. Ultimately, the study noted that some larger issues exist including that most speech-recognition engines are trained on and expect an adult voice instead of a child's voice and that there is, in general, a much larger corpus of adult speech available for training. Although there have been improvements in the technology since these articles, we encountered issues as we developed the Digital Language Shower as well.

Aware of some of the issues raised in the previous research, the L2TOR project [17] had the goal of creating a second language tutor using a social robot for children ages five and up. The project documented several different aspects that should be considered when developing such a solution. These aspects include pedagogical issues such as what words to teach, what ages to target, and how many words to teach. Other aspects were in child–robot interaction such as how to introduce the robot, the interaction context, the robot's nonverbal and verbal behaviors, how to give feedback, and the importance of customizing the lessons to each child.

Then, the L2TOR project applied these guidelines and ran a study that examined how a NAO robot and tablet compared against only a tablet for teaching English words to Dutch-speaking children [30]. The tablet was used as a medium for the robot to communicate with the children to avoid issues with ASR. One-hundred ninety-four Dutch children between the ages of five and seven were split into groups that had seven lessons one-on-one with either a robot and tablet, a tablet only, or no technology. The robot group was further divided into groups where the robot used gestures to pantomime the words and one that did not. The study found that children could retain the words they had learned from the robot at the same level they could with a tablet and that there was no difference between groups where the robot used gestures or did not.

2.3. The Language Shower Concept

The original Language Shower was developed by the pedagogic staff in the Grorud district of the city of Oslo [1] and is based on research from children's language development in school and preschool. It contains of a set of Norwegian words that the children are supposed to know when they start school. The entirety of words is split into topics, such as the human body, my family, home, and clothing. A typical week with Language Shower consists of one daily session lasting 15 min with groups of six to eight children and one or two staff. The program follows a basic two-week schedule: during the first week (five days), roughly ten new words are introduced and learned, and during the second week, these words are frequently repeated and used in various tasks. This pattern is then repeated with different words and topics.

Each word is accompanied by a picture, such that for instance "hand" comes with a picture of a human hand as a visual aid. The staff (both pedagogues and other adults) controls which pictures are shown, explain words and tasks, and evaluate answers and each child's performance. The program also provides different tasks related to the words, such as multiple-choice questions, mind mapping with synonyms and antonyms and other related

terms, comparisons, clapping, and rhymes. The original version of Language Shower started with pictures presented on sheets of paper, but there was also an implementation in PowerPoint where the pictures were shown on a wall by means of a projector connected to a tablet PC.

3. The Digital Language Shower Solution

We transformed the existing Language Shower to a new digital format with a tablet app and an optional robot. We tried to follow the original learning program as closely as possible (with certain modifications to exploit the possibilities that the novel combination of app and robot gives). There were several reasons for this:

1. The original program was revised multiple times to incorporate experiences during daily use in day-care centers.
2. Keeping similarities between the old and the new format would make it easier to identify those factors that are potential causes for differences in learning.
3. We wanted a solution where control remains with the pedagogic staff.
4. Finally, we then could reuse a number of pictures and tasks from the existing Language Shower, giving a more efficient content development.

The entire project started with a small-scale pre-project in autumn 2018 to assess the maturity of the technology [3]. Early in 2019, we conducted interviews and focus groups with (ordinary) staff from the day-care centers, the district's pedagogues and experts for language training, as well as parents, 14 parents in total, both with and without an immigrant background. The notes from the conversations were then structured (by a single researcher), analyzed in order to derive underlying requirements, which were in turn divided into user requirements and functional requirements. The suggestions included to make provision for a dialog between the children and robot and to include game elements to increase the children's motivation. The interviews also reinforced the need to involve the children's parents in the language teaching as the parents are crucial for maintaining the children's motivation at home. We observed through the interviews that many of the parents needed to learn or improve their own Norwegian skills.

The Digital Language Shower consists of an app for smartphones or tablets and a social robot. Both are interconnected. The software development was carried out in two major iterations: A beta version of the app and the robot were tested in a limited pilot in early 2020. The suggestions for enhancements from the trials were implemented along with bug fixes and new features before the main trial in autumn 2020 and complemented with all necessary content. This resulted in the final version.

Previously, we gave a short outline of the solution and its components in a late-breaking report [5]. We now present a detailed explanation of the content, the app, and the robot, starting with the Digital Language Shower requirements.

3.1. Technical and Functional Requirements

Finding the appropriate set of requirements was an involved process since there were multiple target groups:

1. The children in the day-care centers;
2. Their parents and families at home;
3. The pedagogical staff.

In addition, there were security and privacy aspects to take into account, as well as the existing pedagogical framework, as detailed further below.

It was viewed early on as beneficial to combine the functionality of the app with the robot for several reasons. First, there needed to be an additional means to show the visual cues of the words the children were supposed to learn. Second, the children's families would not have a robot at home, but the availability of a smartphone is common in Norway. Third, learning can continue using the app regardless of whether the robot is connected or not (especially in cases of a poor network connection). This also neatly addresses the

principle of *progressive enhancement*, where the app provides the basic learning experience that is further enhanced by the robot if it is available.

Next, there were requirements from the day-care center staff. In practice, unexpected situations can easily arise, such as children talking simultaneously, touching the robot, and children who require extra attention, leading to the requirement of a nonautonomous robot. Last but not least, the app could take over both the tasks of a content server and monitoring of the learning progression, as explained in Section 3. This removed the need for external dedicated servers and thus reduced the solution's complexity, costs, and maintenance efforts quite a bit.

An easy-to-use and cost-effective app appeared to solve all the aforementioned demands.

3.2. Content

This section explains the details of the content, including the learning program and pedagogical concepts. By content, we refer to media files, such as images, sound effects, music, speech, choreographed robot movements, other robot instructions, and descriptions of how these media files are organized, interconnected, used, and presented. Content is both a representation and a description of digital assets for learning and, therefore, fills the app and the robot with "life".

3.2.1. Course Structure and Modalities

The Digital Language Shower program was designed as a course that is subdivided into course modules, also referred to as topics. The topics are, in turn, a compound of lectures. A lecture consists of learning units. The semantic description of modules, topics, lectures, and learning units is stored in one file in JSON format. The content file contains the instructions for both the app and robot, and it points to where the media files for images and audio are stored in the file system. Each semantic element holds a unique ID. The description also specifies the achievable score for each learning unit.

A lecture's maximum score is the sum of scores for its learning units, and a module's maximum score is accordingly the sum of scores of the module's lectures. It is further possible to specify a minimum-score requirement for entering a module or lecture, combined with particular content IDs; otherwise, access to the lecture is locked until the score is achieved. This mechanism controls the order for accessing content and links this to the user's learning progress.

All learning units have in common that the user actively must choose to continue, e.g., by pushing a button. However, depending on the specification of the learning unit, this choice may not be available without an answer, or before, say, a playback button has been pushed.

A learning unit may contain descriptions of alternative content called modalities, which are the most basic logical structures. The learning goal of a set of associated modalities is identical, but the learning means or measures typically differ. In this work, we developed two modalities for the content:

1. App-only;
2. The app in combination with the robot.

The first modality describes the learning means for situations without the robot, such as at home, or when the robot and app are not interconnected, while the second modality enhances the lectures with the robot, i.e., targeting day-care centers.

3.2.2. Learning Unit Types

We developed three basic types of learning units:

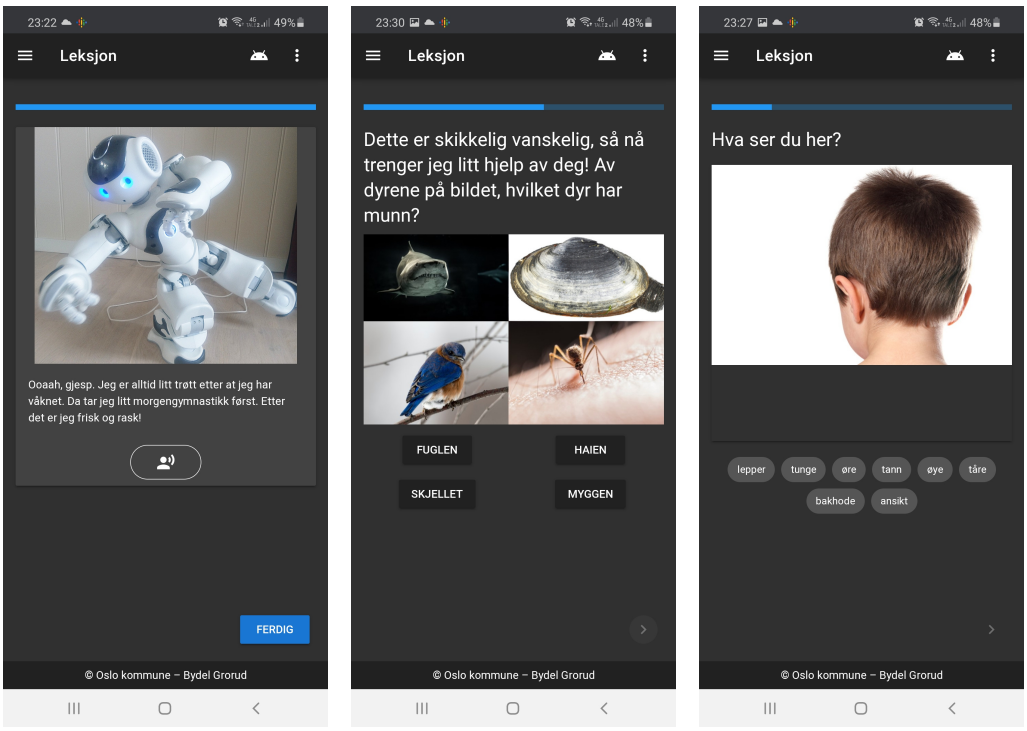
1. A plain narrative (listen or watch);
2. A multiple-choice question;
3. A word collection.

Figure 1 shows how these are displayed in the app. A basic characteristic of the learning units is whether they provide for one-way communication of knowledge or two-way interaction.

Plain narrative units (Figure 1a) imply no interaction. In the app-only modality, a plain narrative unit typically means the display of an image and a piece of text or the playback of an audio file, containing for instance the pronunciation of a particular word. Using the app in conjunction with the robot, the same image is shown in the app, accompanied by the robot telling a story or playing a song and dancing. Regardless of the modality, the key element is the image, which is supposed to trigger the children's associative abilities.

A multiple-choice unit (Figure 1b) depends on interaction, that is somebody answering the question. The app displays one or several images as a visual clue and, in the app modality, a textual question with two or multiple answering options as buttons. In the modality of the app and robot, the app shows only the image, and the robot takes over the dialog, i.e., poses the question and awaits the answer. Here, it is up to the content producer to specify how many times a question can be answered. In our study, we chose three trials for all oral multiple-choice questions. As an answer is given, the robot replies with what it has understood first and whether the answer was correct. Thereafter, as long as it still accepts more answers, it responds "Again". Should the answer still be wrong, the robot says the correct answer and gives a few comforting words. A correct answer is awarded by the robot with a victory position, a little dance, clapping, or another funny move. In the app-only modality, it is only possible to answer once.

A word collection unit (Figure 1c) also relies on user interaction. The app displays a set of images, one at a time, typically with words to be repeatedly trained. In the app-only modality, there also is a set of buttons, one for each word, and the user has to press the right button of the currently raised image. In the app and robot modality, the robot asks for the word of the current image and evaluates the answer. The images shift after each answer, and the entire unit is finished after the last image has been presented. A word collection unit differs from the other units in that each wrong answer leads to a reduction in the score, but the score cannot go below zero.



(**a**) Plain narrative (**b**) Multiple-choice question (**c**) Word collection

Figure 1. Screenshots of tasks/learning units in the app.

3.3. Pedagogical Aspects

We produced content for a six-week learning program in total: two weeks with the topic "Clothes" and four weeks with "My body". In terms of word count, this implied respectively 10, 11, and 12 new words in the word introduction weeks, each of which was followed by a repetition week.

We further extended Language Shower with a couple of new elements: There is a module called "The robot" with three lectures. In the first lecture, the robot says hello, and this lecture can be played the very first time the children meet the robot or when a session is started. In the second lecture, the robot pretends to just have woken up, plays back the opening song, and dances. The opening song is supposed to be played before each Language Shower with the purposes to put the children in the right state of mind. The last lecture contains an ending sequence where the robot waves and says goodbye. This one is supposed to be played at the end of each language-learning session, for the children to more easily accept that the session is over.

When there were not enough tasks in the original Language Shower, we added new content (five to eight tasks), with the goal to fill approximately 15 min for each session. We tried to vary the tasks as much as possible within the framework given by the original Language Shower program. In addition, the robot's moves, lights, audio, and phrases were randomly changed to achieve more variation. For example, the sentence the robot says to comfort students when an answer is incorrect has over six alternatives, which should minimize the effect of the robot repeating itself every single time. Finally, we also added one dance every other week, including appropriate moves and light effects. The music was purchased from professional producers.

A day where a new word was introduced started with a repetition of words learned earlier in the week (except on the first day of each week). Then, two to three new words were introduced, and each new word had a combination of a multiple-choice question ("what do you see here") followed by a task to make a mind map for this new word. The mind map task is implemented as a plain narrative unit and solved by the group while the robot waits. The mind map activity relies heavily on the staff's guidance. Typically, the remaining tasks consist of pronouncing the words' first letter, clapping syllables, and finding rhymes. This is followed by associative tasks regarding color ("red shirt"), count ("three fingers"), size ("large legs"), position ("bottom left"), comparison to related phrases ("yellow like cheese"), and so on.

A repetition day starts with a word collection containing all words introduced during the week before. After this task, the focus turns to two to three selected words from the collection. A typical activity here is to go through the mind map for each word again and expand the map. Following this task, more tasks with associative and interactive activities are performed. The tasks here are typically more difficult than during the introduction week.

3.4. Robot Instructions

The content file also contains robot instructions, either a single instruction or a sequence of instructions, such as:

1. Raise left arm;
2. Play back a sound;
3. Then, say "correct".

As a movement can happen concurrently with audio and speech, almost any robot behavior can be described by an instruction set such as this. During the execution of a learning unit, the app splits a series of instructions and sends the robot single instruction messages. See Section 3.7 for a description of these.

An annotated example of a lecture description is given in Listing 1 with a plain narrative unit that consists of two modalities. The example shows Norwegian texts only, but in general, each piece of text must be specified in all available languages. As speech is always in Norwegian, no further language specification is currently needed.

Listing 1. An example description of a lecture.

```
{
 ID: "f9819e7f-d30d-4e84-adff-ad5c618e37ce", // unique ID
 Title: [
 {
  Language: "no", // Norwegian
  Text: "Ha det" // "See you"
 }
 ],
 Description: [
 {
  Language: "no", // Norwegian
  Text: "Roboten vinker og sier «ha det»." // "The robots waves and says 'see you'."
 }
 ],
 Illustration: "robot-waving.jpg", // image illustrating the lecture
 Requirements: [], // here, no requirements to enter lecture
 Units: [ // list with learning units contained in this lecture
 {
  ID: "d58e46ff-88ec-46d2-ae41-bf0ebe5584d3", // learning unit ID
  Score: 0, // maximum score
  Modalities: [
   { // description for combination of robot and app
   Type: "narrative: robot + still",
   Still: "robot-talking.jpg", // image shown by the app
   Instructions: [ // for the robot
    {
     Type: "monologue", // robot doesn't wait for answers
     Audio: "gameover-magical.mp3" // playback in robot
    },
    {
     Type: "monologue",
     Speech: "Okei, nok for i dag. Da ses vi kanskje en annen gang?" // "OK, that's it for today. See you next time!"
    },
    {
     Type: "monologue",
     Speech: "Ha det bra, ha det bra alle sammen.", // "Good bye, everybody"
     Gesture: "Greeting" // the robot's movement
    },
    {
     Type: "monologue",
     Speech: "Sånn. Da kan jeg gå og sove litt." // "Now I think I'll go and sleep"
    },
    {
     Type: "monologue",
     Audio: "yawn-long.mp3"
    }
   ]
   },
   { // description in case of app only (no robot)
   Type: "narrative: text + still",
   Still: "robot-waving.jpg",
   Text: [
    {
     Language: "no",
     Text: 'Okei, på tide å si "ha det"! Vi ses neste gang ...' // "OK, good bye for now. See you next time!"
    }
   ]
   }
  ] // end of modalities
 }
 ] // end of learning units
}
```

3.5. Mobile App

The municipality required that the app run on the Android and iOS operating systems. It was thus decided to implement the app as a Progressive Web Application (PWA). Thus, the app can run on many smartphones, tablets, and PCs with a variety of screen sizes and orientations. PWAs are also great for rapid development and deployment.

The app consists of three main views (not counting the Settings and About views): list of course modules (topics), module/list of lectures, and lecture/learning units. As for now, there are the three topics: *Meet the teacher*, *Clothes*, and *Your body*. A single lecture corresponds to one Language Shower session, i.e., the children are supposed to go through one lecture each day, or—in turn—the learning units contained in a lecture. An example for a lecture is: "New words: Underpants, boxer", and "Repetition: All previous words, particularly sock, tights". During a lecture, the user navigates from learning unit to learning unit by pressing the Next button in the screen's lower-right corner.

The connection parameters to the robot can be personalized in the settings of the NAO. When the robot's chest button is pressed, it speaks its IP address, which, in turn, can be entered in the respective field in the app's settings. This is followed by an internal handshake between the app and the robot. If the handshake is successful, the robot icon in the app's lecture view will turn green.

The app also provides some gamification elements in the user interface. The lecture and course module scores are shown as the user has finished a lecture and also later in the lecture listings and module listing. Some graphical user interface elements slide in and out of view in an animated, dynamic manner and are accompanied with sound effects. Correct answers produce a bell sound, and incorrect answers produce a buzzer sound.

3.6. Monitoring Progress

After a learning unit is finished, important key parameters that measure the user's progress are collected and stored as a progress report inside the app's internal database in xAPI format by the Advanced Distributed Learning Initiative [31]. These measurements include starting time, unique user ID, unique learning unit ID, ending time, score, and maximum score. Hence, each progress report corresponds to a single learning unit and summarizes who has done which learning unit, when, for how long, and how successful the progress was. The app uses the database to keep track of which lectures a user already has done, and the scores for lectures, modules, and courses so far. As scores already achieved elsewhere can be retrieved, it is possible to control which lectures, modules, and courses should be available to the user based on earlier progress.

In the trials discussed in Section 4, the progress report was also sent to an external database called the Learning Record Store (LRS), which was hosted on our servers so that we could track the progress remotely. As a side note, these progress reports allow for the possibility to track users' learning progress and to analyze the content for potential weaknesses. This *was not* part of the current project.

3.7. Robot

The project used two NAO 6 robots that were semi-autonomous. That is, the robot would receive instructions from the app, perform its actions autonomously, and then wait for the next instructions. When building our version of Language Shower, we had to determine how to best use the robots' speech capabilities and how to control the robot through the app.

3.7.1. Speech Recognition and Speech Pronunciation

From the factory, NAO has the capabilities of speaking and understanding English, but one can install a Norwegian voice and speech recognition software. Given the voice that it uses, we assumed that the voice technology was based on Acapela's system. NAO's Norwegian pronunciation is not prefect, but it has improved from previous versions and is considered usable. Fortunately, it is possible to adjust the pronunciation by adjusting

the text to be pronounced. We tried Acapela's tags for controlling NAO's speech [32], but they did not seem to work. Eventually, we adjusted the regular text, including using accent marks, splitting words, and adding punctuation in creative ways to obtain an acceptable result. For example, the robot pronounced the location name "Grorud" as [groru'-], but it should actually be pronounced [gro:'ru-]. We adjusted the input text to be "Grô Rud", which is close enough. In other situations, we could not get the robot to pronounce the "s" sound correctly, and had to adjust the word accordingly. We recognize that this technique is fragile and could break upon updates to the speech synthesis software [33]. However, this was the best we could do with the current setup and limited time for development.

As mentioned in Section 2.2, we ran into limitations with NAO's built-in Automatic Speech Recognition (ASR), which was lacking sufficient documentation for Norwegian. NAO's Norwegian speech recognition could not understand words that were not in its dictionary, such as names. This limitation ruled out some simple dialogues. For example, the following construction would *not* be possible (emphasis added for word not in the dictionary):

Child: "Hello, Robot. My name is *Suleiman*."

Robot: "Hello *Suleiman*."

This dialog would not work because the name *Suleiman* was not in the dictionary. It may have been possible to add these names, but there was no documentation for this.

Further, the ASR had difficulties distinguishing between simple words. For example, the Norwegian *en* (one) was misinterpreted as *fem* (five); occasionally even clearing one's throat was misinterpreted as "five." We considered lowering the confidence threshold for the recognition, but this even worsened the quality of the recognition.

We addressed this challenge by working with an expert from our project partner Innocom, who implemented a solution using Google's Speech to Text service. This introduced a small delay, as speech was sent to Google's service for processing before the recognized text was returned. Our tests showed that the delay was negligible in most situations. There were also issues with Google's service, which we detail later.

3.7.2. Controlling the Robot and the Language Shower

To address the requirement for progressive improvement of the learning experience, it was decided that a pedagogue would control the Language Shower session through the tablet app. That is, the robot imitated the session leader, but it presented the instructions sent from the tablet app controlled by the pedagogue who actually led the activities. Due to this dependency between the tablet app and speech recognition, the robot always needed to be connected to the network. In our study, we used an internal network at the day-care centers for the robot and the app.

The robot was connected to the app using the WebSocket protocol [34]. After the robot had received its IP-address from the router and this address was added to the settings dialog in the app (Section 3.5), the app would send the robot one instruction at a time. The robot would then send a confirmation that it had completed the instruction before the next instruction was sent. The instructions were formalized in JSON, and the communication served as an API between the robot and the app.

Listing 2 shows a message from the app to the robot, and Listing 3 shows the corresponding response from the robot to the app. The most important fields in Listing 2 are:

1. Speech (Line 10): what the robot will say;
2. Gesture (Line 11): how the robot should move;
3. Light (Line 12): control the lighting effect in the robot's eyes;
4. Color (Line 13): control the color of the robot's eyes.

Listing 2. An instruction from the tablet app to the robot.

```
{
 "Title": "NR Tutor: App", // Message from the app
  "Version": "0.9",
  "AppID": "...",
  "DateTime": "2018-11-13T20:20:39+00:00",
  "DeviceID": "...",
  "Request": "Directive", // Type of instruction
  "ObjectID": "6db6...611e", // Instruction UUID
  "AnswerRequired": false, // No interaction required.
  "Speech": "Hello everyone, I am Robbie", // Words to say.
  "Gesture": "winner", // The gesture to perform.
  "Light": "blink", // Effect for the eyes.
  "Color": "\#ff0000" // The color the eyes should have.
 },
```

Listing 3. Confirmation response from the robot to the app's instruction in Listing 2.

```
 {
  "Title": "NR Tutor: Robot", // Message from the robot
  "Version": "0.9",
  "AppID": "...",
  "DateTime": "2018-11-13T20:20:39+00:00",
  "DeviceID": "...",
  "Response": "Directive",
  "ObjectID": "6db6...611e", // Instruction UUID app sent
  "Done": true // Robot done with the app's instruction
 },
```

The NAO's actions were programmed using Choregraphe [35]. Using the program, we created a content package, installed it on all the NAOs, and configured the package to start automatically. The package contained the code for the robot's actions and the WebSocket client. The robot does not perform any actions until it receives instruction from the app. If no instructions are received after ten minutes, the robot will go into its resting (crouch) position to save the battery and reduce wear on the servos. One could also change the robot into this position by touching the back of the robot's head.

There were some additional challenges with the specific version of NAO we used. NAO runs its own custom version of Linux (based on Gentoo), which could cause issues when adding additional software. Our version of NAO and Choregraphe used Python 2.7 as the interpreter. Although this version of Python was supported when the NAO was originally released, it is no longer supported [36] as of today. This raised issues when we wanted to make the solution more secure and robust, as follows.

One specific issue arose while trying to implement the Transport Layer Security (TLS) for the solution. Since the app and robot communicate together, it is good practice to run the connection over TLS to ensure the robot is talking to the correct app and the app is talking to the correct robot. Unfortunately, it was difficult to add certificates to the robot's certificate store as the required encryption algorithms were missing. As a result, it was not possible to complete a TLS handshake between the robot and the app. Ultimately, the attempt was abandoned due to a lack of time. We instead focused on improving security by running the robot on a closed network.

Another issue came with the older version of the tornado library, where its WebSocket implementation did not handle the case of being unexpectedly disconnected correctly. There were features implemented in later versions of the library, but we could not upgrade as it was used by other parts of the system. It might have been possible to build a custom version of Python with the correct libraries, but this would have added additional complexity in how to communicate between the different Python processes. Again, we had to abandon this due to time limitation and came up with the send-and-confirmation protocol to add additional robustness to the solution.

4. Trials

There were two trials of Digital Language Shower. In both trials, the app was supposed to be tested also by a group of parents. Unfortunately, this proved difficult to achieve due to circumstances related to the COVID-19 pandemic. It was, thus, decided to drop the trials with parents and focus on the day-care centers instead. All trials were registered and approved by the Norwegian Center for Research Data in advance.

4.1. Pilot (with the Beta Prototype)

The beta versions of robot and app were tested in a pilot trial, which was limited to a single group of six to eight children (with a day-to-day variation in size) and two staff in a single day-care center over a period of six days early in 2020. Only the words from a single Language Shower topic were introduced. The recruitment of a small population of children with an immigrant background was organized by the day-care center. The beta version was a proof of concept and, thus, subject to many potential improvements. It is stressed that the focus in the pilot was on getting the app and robot to work seamlessly together and few observations were planned, without a systematic analysis of the learning experience and without quantification of the learning impact as in the main trial.

As outlined in Section 3.5, a Language Shower session lasted approximately 15 min, and its progress was controlled by the pedagogic staff. Researchers observed the session and how the children, staff, app, and robot interacted. They took notes, pictures, and videos and interviewed three of the staff after the children had left. The observation and interview notes were analyzed, structured, and turned into requirements for further improvements of both the robot and app.

4.2. Main Trial

The main trial was conducted with the improved Language Shower over a period of six weeks in two day-care centers in the Grorud district in autumn and winter 2020, covering words from the three Language Shower topics presented in Section 3.5. As with the pilot, the recruitment of children was organized by the day-care center. In order for the participating testing groups to be sufficiently large, we had to accept a high degree of heterogeneity in age (3–6 years), language skills, as well as cultural and other backgrounds. A few days before the actual trial, the staff received a demonstration of how to handle the app and robot, and they were given a detailed instruction set and schedule for the lectures. None of the children had been exposed to Language Shower earlier. Some of the staff had some experience with this program.

To measure the robot's impact in an objective manner, we planned a quasi-experimental design with one Robot Group (RG) and one Control Group (CG) per day-care center. There were four groups in total: two RGs and two CGs. The robot groups should use both the app and robot, while the control group would go through Language Shower in its original way. We had planned for seven children per group, twenty-eight children in total, but this number was eventually reduced to twenty due to illness, no-shows, and other circumstances. For instance, some were kept at home due to the pandemic situation, and others turned out to be simply too shy to participate. All children had an immigrant background. The distribution of children per day-care centeris shown in Table 1 and per group in Table 2. Both genders (eleven female, nine male) were represented.

The children's progress in language skills was measured in a pretest and a post-test. Each test counted the number of words that a child was able to recognize and say aloud, once before the trial and once after. Each participating child was shown thirty-three small cards in total, one at a time, with pictures of words from Language Shower. The entire session was carried out in an informal and playful manner so that the children did not experience the session as a test. The pretest was conducted one week before the trial, whereas the post-test, which originally was planned for the week after the trial, had to be postponed by a few weeks due to the COVID-19 pandemic. The final time between the tests was 10 weeks. The delay affected the groups equally and therefore did not have an

effect on the comparison across groups. One boy showed great progress during the trial and took the pretest in English and the post-test in Norwegian; this was considered as valid data.

Table 1. Distribution of gender and trial groups over day-care centers.

	Girls	Boys	Robot Group	Control Group
Day-care center A	7	5	7	5
Day-care center B	4	4	2	6
Sum	11	9	9	11

Table 2. Distribution of gender between trial groups.

	Girls	Boys	Sum per Group
Robot Group	6	3	9
Control Group	5	6	11
Sum per Gender	11	9	20

There were two further methods of data collection: observations and interviews. We wanted to get a first-hand impression of the new Language Shower, to become familiar with the children's learning experience and the staff's user experience, and to note down any technical difficulties (and solve critical ones). We therefore observed the children in situ. We had 13 observation sessions with all RGs and CGs in both participating day-care centers. Most of the observations were performed during the first and the last week of the trial. We expected excitement during the first week plus some potential technically critical challenges that needed to be solved. By waiting as long as possible between the first and final observation, we anticipated we would capture the biggest difference in the staff and children's interaction pattern. Due to the circumstances of the COVID-19 pandemic, three observations were conducted remotely, i.e., by video conferencing. This reduced the quality of the observation somewhat. The observations were noted down in writing and further enriched with pictures and videos.

Furthermore, we performed interviews with four of the pedagogic staff who conducted Language Shower for both RGs and CGs at the participating day-care centers. The interviews were semi-structured and conducted—due to budget limitations—by a single note-taking researcher, who was guided by a number of predefined questions. Both the notes from the observations and interviews were then subject to a thematic analysis [37], which was carried out in a joint manner by the same researcher—again, due to a tight project budget—to identify recurring topics and threads.

5. Results

In the following, we detail the findings from the different trials.

5.1. Pilot Results

The app was tested and worked satisfactorily under lab conditions. However, at the beginning of the pilot, we had to solve several unexpected connectivity challenges, as connection to and from the robot was blocked by network firewalls at the day-care center. The situation was further complicated by the tablets being configured to connect to multiple networks and automatically hopping to a different network with a stronger signal. As a consequence, the robot and tablets could suddenly be on different networks when signal conditions changed, with the connection between the robot and tablet suddenly blocked by a firewall. The remedy was to restrict the robot and tablets to a single network and open the port in that network's firewall for communication between the tablet and robot.

Subsequently, we discuss selected observations and topics encountered in interviews. An exhaustive listing of all observations and interview quotes is outside the scope of this work. We observed for instance that the staff were confused about how the robot was to be used, and this led to the app's how-to section being extended and partly rewritten. Another example was that the solution was developed using Android tablets, but the day-care centers used iPads; as a result, the playback of audio files in the PWA did not work on iPads. We solved this by re-encoding the audio to a common format for both platforms. It was also observed that the robot often classified answers as wrong, even though they were correct. The remedy here was to incorporate many more possible answers into the list of correct answers and to loosen the checks for correct answers.

The staff, in general, had a critical view towards Language Shower, which at that point was not conducted on a regular basis. In spite of the critical view, they commented that the program appeared to work well and that the robot seemed to be "an exciting element of fun" in the language-learning process. The staff particularly liked how the robot could be controlled by the app.

At the same time, the staff was concerned about how to ensure that answers were marked correctly. The robot would play an audio signal to indicate it was listening for the answer, but it was common that the children (often, many at the same time) would answer the question before this signal was played. This led to many situations were the ASR was not able to detect what had been answered, and consequently, the robot stated the answer was wrong. Another topic was group size. The staff was concerned that a single adult would not be able to control the robot and simultaneously steer eight children aged three to six.

5.2. Main Trial Results

Here, we present the results from the main trail's vocabulary test, observations, and interviews with staff members.

5.2.1. Vocabulary Test

The number of correct words was counted for each child in each test, and the difference between the post-test and the pretest was then calculated, leading to a child's *score* (or Δ). The mean scores per trial group are shown in Table 3, the standard deviation and the score as a percentage of the 33 total words in the program. Six children performed almost equally in the pre- and the post-test, with their scores being approximately zero (± 1), i.e., they did not have any progression. For the majority, however, a learning effect could be observed: On average, the children in the robot group had learned 7.2 more words over the trial period, an improvement of 21.8 %. In the control group, the learning effect was smaller: Here, a child had learned on average 3.1 more words, an improvement of 9.4 %. This makes the robot group's progress approximately 12.4 % larger than the control group's progress.

Since the data did not have a normalized distribution, we used a Mann–Whitney U-test on the mean Δ. The test indicated that the learning effect was greater in the robot group ($\mathrm{Mdn} = 8, n = 9$) than in the control group ($\mathrm{Mdn} = 4, n = 11$), $U = 75, Z = 1.95, p = 0.028$. The effect size $r = 0.436$ indicated a medium effect. We also calculated bootstrapped 95% confidence intervals for the mean Δ to give an idea of how much the differences overlapped.

Table 3. Vocabulary scores per trial group. The 95% confidence interval (CI) was calculated using bootstrapping (* = $p < 0.05$).

	Min	Max	Mean Δ*	95% CI	Std. dev. Δ	As pct. of 33 Words
Robot	−1	13	7.2	(4.1, 9.9)	4.9	21.8%
Control	−1	9	3.1	(1.0, 5.1)	3.5	9.4%

Tables 4 and 5 show the differences between the groups in the same day-care center and differences between genders, respectively. Since the number of participants in each

subgroup was small, we do not provide statistical tests. All groups showed improvement between the pretest and post-test. Table 4 shows that the groups in day-care center A did better on the average than the groups in day-care center B. There was a similar development when comparing between genders (Table 5). There was, on average, an improvement regardless of whether a boy or girl was in the control or robot group, but the participants in the robot group learned on average more words than the control group. The girls learned on average 6.4 more words in the RG versus the CG, while the boys learned on average 2.3 more words in the RG versus the CG.

Table 4. Breakdown of vocabulary scores for day-care centers by group; the 95% confidence interval was calculated by bootstrapping.

	Mean Δ	95% CI	Std. dev. Δ	As pct. of 33 Words
Robot Group A	8.1	(5, 11)	4.3	24.7%
Control Group A	5.2	(2.2, 7.6)	3.4	15.8%
Robot Group B	4.0	(−1, 9)	7.1	12.1%
Control Group B	1.3	(−0.5, 3.5)	2.7	4.0%

Table 5. Breakdown of vocabulary scores for genders by group; the 95% confidence interval was calculated by bootstrapping.

	Mean Δ	95% CI	Std. dev. Δ	As pct. of 33 Words
Robot Group Boys	7.3	(−1, 12)	7.2	22.2%
Control Group Boys	5	(3.3, 6.8)	3.4	15.2%
Robot Group Girls	7.2	(4, 10)	4.2	21.7%
Control Group Girls	0.8	(−1, 4)	2.7	2.4%

5.2.2. Observations and Pedagogic Staff Interviews

We present the findings from the interviews combined with our own observations in a combined manner and according to the identified themes from our analysis. These are: setup, language learning (concept), robot, technical issues, and gamification.

Setup: A couple of comments considered the Language Shower program itself, i.e., without a robot. Staff from both day-care centers commented that a single adult running the sessions was, in most cases, not sufficient, particularly with children requiring much attention, and in case the technology failed. While the number of children per group varied between five and eight during the trials, several staff members considered six to be the maximum group size.

There were multiple opinions regarding the children's age. The children that participated were aged three to six. One staff member believed the tasks to be too difficult for the youngest children. Another thought the words were too easy for the oldest children. A third informant thought it would be better to deploy Language Shower early in school.

Several felt that the groups had been too heterogeneous in their knowledge of Norwegian in the trials, and groups with a similar level of Norwegian would be beneficial for the success of Language Shower. Some children had dropped out in both the RG and CG. The staff's explanation for this was that the most skilled children had quickly lost interest, while a few children were afraid of answering incorrectly.

Language learning concept: Obviously, there were different preferences regarding language-learning programs among the staff. One staff member recommended copying elements from other programs, for instance the program *Sproff* [38], which had been used earlier in day-care centers. Other staff had experienced that the children "more quickly become uneasy" with Sproff, and that the "learning factor with Language Shower was higher". Yet another informant expressed their dislike of Language Shower and said that it had little impact and was not directly related to the center's other language-teaching efforts.

In particular, the repetition part was viewed as being "a bit boring", and mind maps were considered to be unsuitable for the youngest. This contradicted the opinion of another staff member who pointed out that the possibility to go in-depth with every single picture in Language Shower—that is, making each task a mind map—was one of the most important properties of the program. A clearly positive impact of Language Shower was mentioned: a girl had been observed to talk exclusively during the sessions, but otherwise remained mute. One staff member expressed indifference towards language-learning programs and said "surely there will come something new soon." Approximately half of the interviewees had intentions to continue with Language Shower (with or without the robot). Last but not least, the group in general needed more time to discuss after the robot had asked a question.

Robot: There were mixed, but mostly negative reactions to Language Shower with a robot. On the positive side, the sessions had been "very exciting" and "fun to have been a part of." Our observations showed that the robot was very attractive for the children in the beginning ("Robbie is cool"), but the excitement diminished as technical problems prevailed ("this one's strange", "stupid robot"). The majority of the staff were surprised in a positive way about the children's engagement with the robot, as they had expected more chaos and distractions. This can be interpreted as a sign of the children's larger engagement and excitement when a robot is present. However, it was pointed out that this might vary greatly: for children with behavioral difficulties, a robot was not expected to solve all attention and motivation issues. According to the staff, songs and dances were the most popular part of Language Shower with the robot, and this corresponds to our own observations. One child was even so enthusiastic about the dances that she wished for a NAO robot as a Christmas present.

Technical issues: Several technical difficulties were mentioned, and several interviewees felt the trial had been a "varied experience". For instance, the technology had not always worked as expected: the robot would sometimes "hang up" or "freeze", and it would not react when talked to. Occasionally it would reject answers as wrong with only slight variations in the pronunciation and with perfectly valid speech by native speakers. This had, not surprisingly, a negative effect on those who were still in the early process of learning the correct pronunciation. We were also told that the gap between what was said and what the robot understood in a few circumstances could be substantial, for instance, "hand" became "kebab", or "big toe" became "football". In rare situations, we observed that the app and robot were simply not in sync.

In other situations, learning units had been misunderstood by the staff. Sometimes, a staff member thought the robot was waiting for an answer, while in fact, it was not, or the robot assumed that an answer had been given and went on to the next task, while, in fact, the group was still discussing or a child was just talking quite slowly. This could lead to confusion later on regarding whether the robot was showing the correct picture or would show another "strange behavior". According to the staff, recognizing speech correctly required that the children were sitting very close in front of the robot, which could be challenging when a larger group of children is sitting around it. All these problems led to a negative impression of the staff, and one informant stated that the robot would "not be missed by anybody in the future".

Gamification: The app's gamification elements such as learning scores and sound effects had an effect on the children and staff. They would mimic the sounds and comment on the scores.

6. Discussion

Let us discuss the findings from Section 5 and see what recommendations we can make from the results and our experience.

6.1. Pilot

The pilot showed that the digital version of Language Shower appeared to fit in to the language learning work of the day-care centers. The quiz tasks and gamification elements in the app also appeared to increase the children's engagement. For instance, both staff and children appeared to be thrilled to have "a robot to play with", as one of the staff put it.

The pilot had also revealed several areas of improvement in particular on the technical side. First, it was critical to get the technology and the components up and running and communicating with each other. Speech recognition struggled when children spoke at the same time and when the one answering was too far away to be heard. Even at the highest sensitivity of the robot's microphones, the robot would sometimes not react at all to the speech. To our knowledge, this issue cannot be solved by adjusting the robot alone. Therefore, the strategy was to instruct the staff that only a single child could answer a given question and that each child needed to orient himself/herself so that he/she faced the front of the robot's head. We also added a functionality for multiple or repeated (typically three) answers, which could be configured in the content file. Overall, the observations and interviews led to 81 major and minor improvements in the functionality, content, user interaction, and user interface before the main trial.

6.2. Main Trial

There was much diversity in the main trial results. On the positive side, the quantitative results from the vocabulary tests tell us that there was a noticeable positive impact of using a social robot in the language-learning process. As observers, we witnessed much engagement and excitement among the children. On the negative side, technical difficulties appeared to reduce the learning outcome and the user experience for the children and staff. The problems with the technology were considered so severe that the district authorities decided, among other reasons, to discontinue Language Shower with a robot. The interviews supported the same conclusions as we found in our observations.

It appears that the pedagogic staff are a key factor for the success of the language sessions. As long as they are not convinced by the program's potential for the children's language development, it is of little use that the program comes nicely wrapped in the form of an attractive app and a cute, cool robot. Another important factor is the staff's ability to engage as teachers, with or without a robot. For instance, we witnessed a pedagogue who was able to motivate and include all children equally as she carried out Language Shower without the robot; all children in the group seemed to have a great learning experience.

6.3. Recommendations

From the observations and interviews with the staff, we present the following recommendations that would need to be addressed in the future. We present the recommendations ordered by theme in Table 6.

6.4. Benefits, Disadvantages, and the Embodiment Effect

The trials uncovered the benefits and disadvantages of the Digital Language Shower program. Some of the critique points to technical matters, whereas other shortcomings can be linked to the original language-learning program. Yet other issues are related to the staff and their acceptance of Language Shower per se.

Most importantly, the robot resulted in a 12.4 % increase in language vocabulary for the children in the robot group; a consistent effect seen in both day-care centers. The progress, however, was unevenly distributed. Some children showed no progress at all, or even negative progress. The latter can be attributed to the natural fluctuation of the experiment, while the former can have many complex explanations, such as differences in age, maturity, cultural background, family constellations, in how much the children's parents support language learning at home, as well as others, including combinations thereof. Unfortunately, the qualitative nature of the experiment did not provide the opportunity to control these parameters.

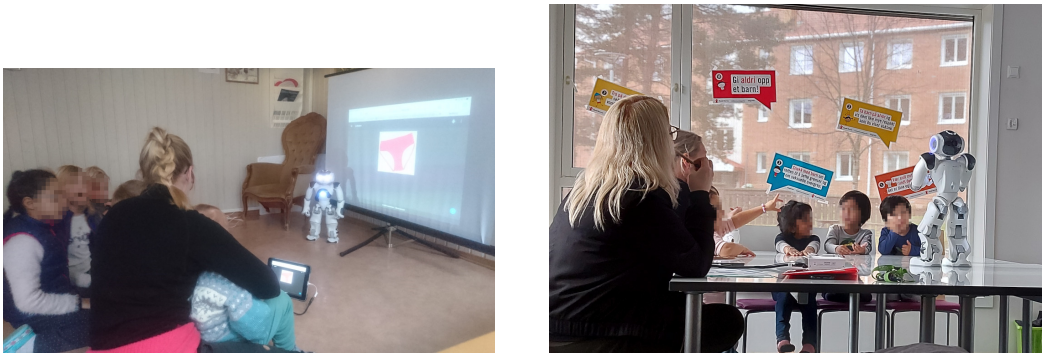
Table 6. Summary of recommendations by theme.

Theme	Recommendation
Setup	• The tablets' screens should always be connected to a projector, such that pictures and tasks are large enough/easy to see. • There should be as few distractions around the group as possible. • It should be mandatory for the pedagogic staff to learn how the robot can be used. • The group size should be 4–6 children. • Groups should be more homogeneous with regard to skill level. • An age of four is considered optimal.
Language learning	• The group should have more time to discuss after the robot has asked a question. • The repetition of words should be less monotonous. • It should be considered to differentiate the learning program with multiple levels of difficulty (for heterogeneous groups and various ages). • It should be considered to employ learning elements from different language programs, such as Sproff.
Robot	• There should be a number of robot dances and songs. • A robot dance should not last longer than roughly one minute. • There should be much variation in task type, robot moves (dances and gestures), sounds, standard answers, etc., to avoid children becoming bored. • The robot should be small, non-threatening, and preferably "cute". • The robot should be particularly careful with actions or speech that could potentially have a demotivating impact.
Technical issues	• Negative learning experiences should be avoided in order not to reduce the effect of embodiment and fascination for technology and robots. • The robot and app should implement robust techniques to sync with each other before and after each learning unit. • There should be as little latency as possible after an answer is given before the robot reacts, but at the same time, the robot should not interrupt in the middle of an answer. • There should be one or several coping strategies for situations when the robot has to wait a long time for an answer, such as "Have you answered yet?". • The speech recognition should become more robust to detect even small differences in pronunciation. The staff should also be able to override the speech recognition result in this case. • The speech recognition should be flexible regarding wrongly pronounced words. For instance, many children replace the difficult-to-pronounce Norwegian "r" with an "l", so they said "plikker" instead of "prikker" (in English: dots).
Gamification	• There should be a number of gamification elements that are reasonably implemented, such that they can increase the learners' motivation while not distracting them too much with graphical and sound effects. • More effort should be put into the proper choice of appealing pictures and songs.

Our observations and measurements confirmed the embodiment effect [39] social robots have on their surroundings. The effect can be observed with both children and adults as the combination of technology, the semi-autonomous behavior, and the human-/child-

like shape, and it manifests itself as fascination and increased attention and engagement, which in turn leads to improved learning. Music and dance were not used so much for learning in Language Shower, but more for entertainment and engagement. Yet, it might be this mix of entertainment and education that made the idea of enhancing Language Shower with a robot so appealing. The embodiment effect's objective influence on the learning progress is reflected by the numbers from the vocabulary tests at the group level; however, as observations and interviews revealed, the fascination may diminish because of the continual technical difficulties. Hence, there are limits to how engaging a robot can be, and future evaluations of the embodiment factor should therefore incorporate measurements of the user experience to reflect overall satisfaction with the technology and how the robot is perceived and accepted. There is some evidence for the limits of embodiment, as some children soon lost interest or even dropped out of the robot group, which, as discussed previously, can be due to multiple factors, such as the lack of a child's maturity, avoidance of embarrassing situations when answering incorrectly, and too little variation in the robot's behavior.

There are a few other aspects related to embodiment: With its height of about 58 cm and a "cute" look, none of the children we observed perceived NAO as a threat. However, we know from related work with autistic children that a humanoid robot of that size could trigger serious reactions of fright [40]. Drawing a parallel to other phobias of, say, mice, it might not be the robot's dimensions that trigger this condition. In our opinion, it matters that the robot is small and comparable to the dimensions of an animal. As such, it can also be placed on a table, and the children can sit around it, as depicted in Figure 2b. Moreover, for both Oslo municipality and the staff in the day-care centers, it was important that the robot appeared genderless (i.e., the robot presented neither as a boy nor a girl). It was therefore an advantage that the NAO 6 was available in gray only, i.e., without potentially "gendered colors", such as blue or pink. In addition, the staff chose the name "Robbie" for the robot since they felt that this name could apply equally well for a boy or girl.



(**a**) Pilot (**b**) Main trial

Figure 2. In situ stills from the trials.

As we observed in the trials, the embodiment factor can, to a certain extent, make up for or draw attention away from the weaknesses in the pedagogical concept, particularly at the beginning of a course. However, there are limits to how much it can cover these weaknesses. For instance, if a mind map is not suitable for the youngest children, then neither a robot giving the task nor one helping with complementing the mind map will solve the situation. Such weaknesses will persist, and a robot might even contribute to the imperfections with its own peculiarities. A suitable example of the latter is our observation that the majority of children quickly became bored when there was too little variation in the robot's responses or music and dances, which also illustrates another boundary of the embodiment factor.

Preschool children, i.e., those aged three to six, vary widely in their maturity and capabilities, including language learning. During the main trial, there were up to eight

children with potentially wide-ranging abilities per group. This complicates finding an adequate level of difficulty in the learning material. We argue that personalized learning for each child in the group by means of a robot might be intricate to achieve, as the proper handling of several individuals at the same time requires sophisticated and flexible programming of the robot's logic. It might, therefore, be the best option with today's state of technology to deploy the robot in a single one-to-one setting. An alternative could be to offer series of tasks (within a single lecture) with increasing difficulty, such that the learning effect is highest at the beginning for the youngest and highest at the end for the oldest. Yet another possibility is to offer multiple lectures with multiple difficulties for multiple ages and allow for heterogeneous groups.

Given the issues we highlighted in Section 2.2 and our context of language learning with a small group of children and a teacher or pedagogue, it appears the technology is (still) insufficient to handle all issues. That is, a robot requires the proper detection and potentially identification of multiple speakers, as well as the handling of simultaneous speech, all of which are not easily (nor inexpensively) solved as of today. Further, language learning, which often relies on small differences in pronunciation, demands for robust speech recognition with support for imperfect pronunciation, language accents, and varied dialects, as well as recognizing new or made-up words. To our knowledge, there are no good solutions for these challenges with today's technology. Solving these issues is particularly important in the context of small (and insecure) children. As illustrated in our trials, some of the insecure children dropped out of the robot group due to their fear of answering the "wrong" way; they may have answered correctly, but were not understood by the robot.

Finally, latency in transmission and processing of data might add to the confusion of children and adults, as they then may be unable to determine whether or not the robot is waiting for an answer, or that the robot has received the spoken text and is instead waiting for a response from the cloud-based ASR service. The robot is supposed to make a sound when it is "listening", but sometimes, this sound was delayed, not played at all, or even at the wrong times. We suspect that problem was related to a bug in the robot's software.

7. Limitations

The findings in this study are limited by the following considerations.

The trials were bound to the tight project budget, with a total of three day-care centers, thirty-four participating children, valid vocabulary test data from twenty children, and interviews with, in total, seven staff. It could easily be argued that these numbers are not representative of children in Norway that come from a non-Norwegian-speaking or an immigrant background. In a research project such as this, however, it is virtually impossible to control the population of children and parents to be representative in an experimental design as ours, while keeping all other factors of comparison equal.

The findings were affected by multiple and subtle factors. For instance, different rooms and surroundings, time of the day for the session, influence of the observers, as well as different interpretations of answers may have impacted each individual result of the vocabulary test. We cannot exclude the possibility that a particular child during the second test had become more confident when interacting with the researcher, leading to a positive impact on the result. It may further be assumed that each child would develop language skills naturally as time passes without having been exposed to Language Shower. This factor would, however, have affected both the robot group and the control group in the same manner.

A further limiting factor is that the analysis of the observations and interviews was conducted by a single researcher. For thematic analyses, it is generally recommended to employ parallel coding of multiple researchers to increase the themes' reliability [37]. It has been argued, though, that the thorough effort of a single, experienced researcher may still lead to a good result [41]. Furthermore, given the tight project budget and the considerable number of observations and interviews, this was not an option for us.

As mentioned in Section 2, many studies in learning language do not take into account the novelty effect. While our study did last longer than some of the studies from Section 2, we did not explicitly work on mitigating the novelty effect. It could be that children in the day-care centers from the study would have also grown weary of the robot. A longer study would be need to be run to know if this was the case or not.

Finally, it is a weakness of the study that the effect of the developed solution on the children's parents could not have been sufficiently studied. Given the situation with the COVID-19 pandemic, we were satisfied to have completed the trials with meaningful results. It has been difficult and time-consuming to come in contact with and consistently involve this important target group in the trials over time.

8. Conclusions

In this case study, we described how social robots can be used to strengthen children's training of a foreign language in day-care centers. Language Shower was transformed into a mobile app with the option for additional enhancement with a humanoid robot to lead lectures. The solution was tested in two major trials over the duration of multiple weeks in several day-care centers in the Grorud district in Oslo municipality in Norway.

Our measurements indicated a substantial positive impact by using the social robot as compared to the same course without a robot. Our observations in situ and interviews with day-care center staff at the same time disclosed a number of challenges with the educational concept, as well as the underlying technology. Ultimately, the combination of these issues led to the district not continuing the use of Language Shower with the robot.

Future efforts of working with incorporating a robot with language learning must consider ASR. ASR needs to be improved and be made more flexible and robust. Other areas to explore should be recognizing and handling simultaneous answers, which is linked to multiple-speaker detection.

Our study met several diverse, intricate issues spanning from obstacles imposed by the COVID-19 pandemic to technical issues with the robot and ASR. Issues with imperfect or failing technology play an important role in adopting new technology, and this (along with other factors) led to discontinuation in our case. Using social robots, our study showed us that synchronization issues and problems with ASR must be improved before performing new trials with a similar setup in other settings. To address other intricate issues, such as the attitudes of the staff, the behavior of the children, and the impact of the COVID-19 pandemic, the research design must take into account such obstacles, so that the negative effects can be mitigated as much as possible. Beyond technology, good cooperation with the staff of the day-care centers and the district leadership is important to ensure a project's success. Despite the obstacles we encountered during the study, we still are convinced that there is a role for social robots in language learning in day-care centers and that this should be investigated further. The positive learning outcome measured from our sample supports this view.

Author Contributions: The authors contributed to this article as follows. Conceptualization: I.S. and T.H.; methodology: I.S.; software: T.H. and T.S.; validation: W.L. and T.H.; investigation: T.H. and T.S.; data curation: T.H.; writing—original draft preparation: T.H., W.L., and T.S.; writing—review and editing: W.L., T.S., and T.H.; supervision: W.L.; project administration: I.S. and T.H.; funding acquisition: I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported through funding by the Delprogram oppvekst og utdanning of Oslo municipality, district Grorud.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the the Norwegian Centre for Research Data (NSD) in advance, Reference Number 331729, dated 29 July 2019.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study, respectively their parents.

Acknowledgments: We thank the pedagogues from Oslo municipality's district Grorud for advice regarding Language Shower, Margareth Sandvik at OsloMet for advice with the vocabulary tests, the day-care center staff for their participation and comments, and in particular, the administrative staff in the district for help with setting up and carrying out the trials. We wish to also thank Kristin Skeide Fuglerud for help in securing the project, method suggestions, and many other seemingly small things that helped during the project. Our project partner Innocom AS delivered the NAO 6 robots used in the project and contributed with technical expertise. We are also grateful to our summer students for helping with content production and vocabulary tests. Last but not least, we thank all parents that allowed their children to participate in this study.

Conflicts of Interest: The authors declare no conflict of interest.The funding organization had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; nor in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
ASR	Automatic Speech Recognition
CG	Control Group
HRI	Human–Robot Interaction
ICT	Information and Communication Technologies
IP	Internet Protocol
JSON	JavaScript Object Notation
LRS	Learning Record Store
PWA	Progressive Web Application
RALL	Robot-Assisted Language Learning
RG	Robot Group
TLS	Transport Layer Security
TTS	Text To Speech

References

1. Fantoft, S. Hvert ord teller [Every Word Counts]. In *Områdesatsingene i Oslo: Årsmelding 2019*; Annual Report; Oslo Kommune: Oslo, Norway, 2019; pp. 56–59. https://www.regjeringen.no/contentassets/012071520f7347a48f8e0e5aa489ab60/arsmelding2019oslo.pdf (accessed on 18 November 2021). (In Norwegian).
2. Breazeal, C.; Dautenhahn, K.; Kanda, T. Social Robotics. In *Springer Handbook of Robotics*; Siciliano, B., Khatib, O., Eds.; Springer-Verlag: Berlin, Germany, 2016; pp. 1935–1972. [CrossRef]
3. Fuglerud, K.S.; Solheim, I. The Use of Social Robots for Supporting Language Training of Children. In *Studies in Health Technology and Informatics*; IOS Press: Amsterdam, The Netherlands, 2018; Volume 256, pp. 401–408. [CrossRef]
4. Gouaillier, D.; Hugel, V.; Blazevic, P.; Kilner, C.; Monceaux, J.; Lafourcade, P.; Marnier, B.; Serre, J.; Maisonnier, B. Mechatronic Design of NAO Humanoid. In Proceedings of the 2009 IEEE International Conference on Robotics and Automation, Kobe, Japan, 12–17 May 2009; pp. 769–774. [CrossRef]
5. Schulz, T.; Halbach, T.; Solheim, I. Using Social Robots to Teach Language Skills to Immigrant Children in an Oslo City District. In Proceedings of the Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, UK, 24–26 March 2020; Association for Computing Machinery: Cambridge, UK, 2020; pp. 442–444. [CrossRef]
6. Bartneck, C.; Forlizzi, J. A Design-Centred Framework for Social Human-Robot Interaction. In Proceedings of the RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759), Kurashiki, Japan, 22 September 2004; IEEE: Kurashiki, Japan, 2004; pp. 591–594. [CrossRef]
7. Cooney, M.; Leister, W. Using the Engagement Profile to Design an Engaging Robotic Teaching Assistant for Students. *Robotics* **2019**, *8*, 21. [CrossRef]
8. Alemi, M.; Meghdari, A.; Ghazisaedy, M. Employing Humanoid Robots for Teaching English Language in Iranian Junior High-Schools. *Int. J. Hum. Robot.* **2014**, *11*, 1450022. [CrossRef]
9. Yamazaki, R.; Nishio, S.; Ogawa, K.; Ishiguro, H.; Matsumura, K.; Koda, K.; Fujinami, T. How Does Telenoid Affect the Communication between Children in Classroom Setting? In Proceedings of the CHI '12 Extended Abstracts on Human Factors in Computing Systems, Austin, Texas, USA, 5–10 May 2012; ACM: New York, NY, USA, 2012; pp. 351–366. [CrossRef]
10. Alves-Oliveira, P.; Sequeira, P.; Melo, F.S.; Castellano, G.; Paiva, A. Empathic Robot for Group Learning: A Field Study. *ACM Trans. Hum.-Robot Interact.* **2019**, *8*, 1–34. [CrossRef]

11. Fels, D.; Waalen, J.; Zhai, S.; Weiss, P. Telepresence under Exceptional Circumstances: Enriching the Connection to School for Sick Children. In Proceedings of the IFIP INTERACT01: Human-Computer Interaction, 2001, Tokyo, Japan, 9–13 July 2001; pp. 617–624.
12. Børsting, J.; Culén, A.L. A Robot Avatar: Easier Access to Education and Reduction in Isolation? In Proceedings of the International Conference on E-Health 2016. IADIS, Madeira, Portugal, 1–4 July 2016; pp. 34–44.
13. Shiomi, M.; Kanda, T.; Howley, I.; Hayashi, K.; Hagita, N. Can a Social Robot Stimulate Science Curiosity in Classrooms? *Int. J. Soc. Robot.* **2015**, *7*, 641–652. [CrossRef]
14. Köse, H.; Uluer, P.; Akalın, N.; Yorgancı, R.; Özkul, A.; Ince, G. The Effect of Embodiment in Sign Language Tutoring with Assistive Humanoid Robots. *Int. J. Soc. Robot.* **2015**, *7*, 537–548. [CrossRef]
15. Kanda, T.; Hirano, T.; Eaton, D.; Ishiguro, H. Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial. *Hum.-Comput. Interact.* **2004**, *19*, 61–84. [CrossRef]
16. Kanda, T.; Ishiguro, H. Communication Robots for Elementary Schools. In Proceedings of the AISB'05 Symposium Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction (Hatfield Hertfordshire), Hatfeild, UK, 14–15 April 2005; pp. 54–63.
17. Belpaeme, T.; Vogt, P.; van den Berghe, R.; Bergmann, K.; Göksun, T.; de Haas, M.; Kanero, J.; Kennedy, J.; Küntay, A.C.; Oudgenoeg-Paz, O.; et al. Guidelines for Designing Social Robots as Second Language Tutors. *Int. J. Soc. Robot.* **2018**, *10*, 325–341. [CrossRef] [PubMed]
18. Kanero, J.; Geçkin, V.; Oranç, C.; Mamus, E.; Köntay, A.C.; Göksun, T. Social Robots for Early Language Learning: Current Evidence and Future Directions. *Child Dev. Perspect.* **2018**, *12*, 146–151. [CrossRef]
19. Schodde, T.; Hoffmann, L.; Stange, S.; Kopp, S. Adapt, Explain, Engage—A Study on How Social Robots Can Scaffold Second-Language Learning of Children. *ACM Trans. Hum.-Robot Interact. (THRI)* **2019**, *9*, 1–27. [CrossRef]
20. Riek, L.D. Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *J. Hum.-Robot Interact.* **2012**, *1*, 119–136. [CrossRef]
21. Randall, N. A Survey of Robot-Assisted Language Learning (RALL). *ACM Trans. Hum.-Robot Interact. (THRI)* **2019**, *9*, 1–36. [CrossRef]
22. van den Berghe, R.; Verhagen, J.; Oudgenoeg-Paz, O.; van der Ven, S.; Leseman, P. Social Robots for Language Learning: A Review. *Rev. Educ. Res.* **2019**, *89*, 259–295. [CrossRef]
23. Alves-Oliveira, P.; Sequeira, P.; Paiva, A. The Role That an Educational Robot Plays. In Proceedings of the 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), New York, NY, USA, 26–31 August 2016; pp. 817–822.
24. Zaga, C.; Lohse, M.; Truong, K.P.; Evers, V. The Effect of a Robot's Social Character on Children Task Engagement: Peer Versus Tutor. In Proceedings of the 7th International Conference on Social Robotics, ICSR 2015, Paris, France, 26–30 October 2015; Lecture Notes in Computer Science; Tapus, A., André, E., Martin, J.C., Ferland, F., Ammi, M., Eds.; Springer: Berlin, Germany, 2015; pp. 704–713. [CrossRef]
25. Chang, C.W.; Lee, J.H.; Chao, P.Y.; Wang, C.Y.; Chen, G.D. Exploring the Possibility of Using Humanoid Robots as Instructional Tools for Teaching a Second Language in Primary School. *Educ. Technol. Soc.* **2010**, *13*, 13–24.
26. Goetz, J.; Kiesler, S.; Powers, A. Matching Robot Appearance and Behavior to Tasks to Improve Human-Robot Cooperation. In Proceedings of the 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003, Millbrae, CA, USA, 2 November 2003; pp. 55–60.
27. Kennedy, J.; Baxter, P.; Belpaeme, T. The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. In Proceedings of the 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Portland, OR, USA, 2–5 March 2015; pp. 67–74.
28. Baxter, P.; Ashurst, E.; Read, R.; Kennedy, J.; Belpaeme, T. Robot Education Peers in a Situated Primary School Study: Personalisation Promotes Child Learning. *PLoS ONE* **2017**, *12*, e0178126. [CrossRef] [PubMed]
29. Kennedy, J.; Lemaignan, S.; Montassier, C.; Lavalade, P.; Irfan, B.; Papadopoulos, F.; Senft, E.; Belpaeme, T. Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction—HRI '17; Vienna Austria, 6–9 March 2017; ACM Press: Vienna, Austria, 2017; pp. 82–90. [CrossRef]
30. Vogt, P.; van den Berghe, R.; de Haas, M.; Hoffman, L.; Kanero, J.; Mamus, E.; Montanier, J.; Oranç, C.; Oudgenoeg-Paz, O.; García, D.H.; et al. Second Language Tutoring Using Social Robots: A Large-Scale Study. In Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Daegu, Korea, 11–14 March 2019; pp. 497–505. [CrossRef]
31. The Advanced Distributed Learning Initiative. Experience API (xAPI) Standard. 2020. Available online: https://adlnet.gov/projects/xapi/ (accessed on 27 August 2021).
32. Acapela Group. Text Tag Documentation: Acapela TTS For Mobile. 2016. Available online: http://doc.aldebaran.com/2-4/_downloads/audio_system_acapelamobilitytexttags.pdf (accessed on 27 August 2021).
33. Apple, Inc. *Speech Synthesis Programming Guide*; Apple Inc., Cupertino, CA, USA, 2006.
34. Fette, I.; Melnikov, A. The WebSocket Protocol. RFC 6455, RFC Editor. 2011. Available online: http://www.rfc-editor.org/rfc/rfc6455.txt (accessed on 27 August 2021).

35. Pot, E.; Monceaux, J.; Gelin, R.; Maisonnier, B. Choregraphe: A Graphical Tool for Humanoid Robot Programming. In Proceedings of the RO-MAN 2009—The 18th IEEE International Symposium on Robot and Human Interactive Communication, Toyama, Japan, 27 September–2 October 2009; pp. 46–51. [CrossRef]
36. Petersen, B. PEP 373—Python 2.7 Release Schedule. 2016. Available online: https://legacy.python.org/dev/peps/pep-0373/ (accessed on 27 August 2021).
37. Guest, G.; MacQueen, K.M.; Namey, E.E. *Applied Thematic Analysis*; Sage Publications: Thousand Oaks, CA, USA, 2011.
38. Årsberetning 2020 [Annual Report 2020]. Utdanningsetaten, Oslo kommune [Dept. of education, Municipality of Oslo]. 2020. Available online: https://www.oslo.kommune.no/etater-foretak-og-ombud/utdanningsetaten/arsberetning-2020/?del=4-3 (accessed on 27 August 2021). (In Norwegian).
39. Deng, E.; Mutlu, B.; Mataric, M. Embodiment in socially interactive robots. *arXiv* **2019**, arXiv:1912.00312.
40. Kumazaki, H.; Muramatsu, T.; Yoshikawa, Y.; Matsumoto, Y.; Ishiguro, H.; Kikuchi, M.; Sumiyoshi, T.; Mimura, M. Optimal Robot for Intervention for Individuals with Autism Spectrum Disorders. *Psychiatry Clin. Neurosci.* **2020**, *74*, 581–586. [CrossRef] [PubMed]
41. Harding, T.; Whitehead, D. Analysing data in qualitative research. In *Nursing & Midwifery Research: Methods and Appraisal for Evidence-Based Practice*, 5th ed.; Chapter 8; Elsevier: Chatswood, NSW, Australia, 2016; pp. 141–160.