Second Language Tutoring using Social Robots

Project No. 688014

L2TOR
Second Language Tutoring using Social Robots

Grant Agreement Type: Collaborative Project
Grant Agreement Number: 688014

D6.3: Output module for storytelling domain

Due Date: **30/09/2018**
Submission Date: **25/12/2018**

Start date of project: **01/01/2016**

Duration: **36 months**

Organisation name of lead contractor for this deliverable: **Tilburg University**

Responsible Person: **Emiel Krahmer**

Revision: **1.0**

Project co-funded by the European Commission within the H2020 Framework Programme		
Dissemination Level		
PU	Public	**PU**
PP	Restricted to other programme participants (including the Commission Service)	
RE	Restricted to a group specified by the consortium (including the Commission Service)	
CO	Confidential, only for members of the consortium (including the Commission Service)	

Contents

Executive Summary 3

Principal Contributors 4

Revision History 4

1 Introduction 5

2 Large-scale study and the role of output generation 5

2.1 Set-up of the experiment . . . 6

2.2 Design of the gestures . . . 7

2.3 Role of the tablet . . . 8

3 Encouraging production of spatial concepts in an L2 8

4 Comprehensibility of recorded gestures 8

5 Conclusion and Future Directions 10

Executive Summary

This deliverable describes the evaluation of the output module as part of the recent large-scale evaluation study. We focus specifically on investigating why there were no significant positive contributions of the robot's use of iconic gestures. In addition, we describe two studies that were conducted in order to further explore the robot's output capabilities. The findings from these experiments, combined with the results of the large-scale study, provide several points of discussion for the human-robot interaction community regarding the design and implementation of the robot's (social) behaviors, as well as inspiration for further improvement of the Intelligent Tutoring System that has been developed within the L2TOR project.

Principal Contributors

The main authors of this deliverable are as follows:

Jan de Wit, Tilburg University
Bram Willemsen, Tilburg University
Mirjam de Haas, Tilburg University
Emiel Krahmer, Tilburg University
Paul Vogt, Tilburg University

Christopher D. Wallbridge, Plymouth University

Revision History

Version 1.0 (25-12-2018)
First version.

1 Introduction

This document describes the development of the output module for the storytelling domain. As mentioned in previous deliverables, no additional lesson content or technical features were added to the L2TOR Intelligent Tutoring System (ITS) specifically for the context of storytelling. However, there is an overarching storytelling component that is present throughout the lessons, which is implemented in the way the lesson content is presented by the robot and the tablet. For example, in a previous study [1] the English target words were introduced by means of a game of *I spy with my little eye*, consisting of thirty rounds of picking the correct answer from a number of items on the tablet screen, with little variation in the events happening on-screen, nor in the robot's speech output. In the recent large-scale study, each lesson had its own theme (e.g., a playground or a zoo) and there was a variety of interactions to support learning, including:

- Touching an object on the screen;
- Moving an object to a desired location on the screen;
- Moving an object to collide with another object;
- Repeating a word or sentence after the robot;
- Performing a physical act (acting out a certain activity, showing left or right hand.

The storyboards for these lessons, which include the details and timing of these different interaction types, are presented as part of Deliverable 2.3. Because the technical implementation of the Intelligent Tutoring System (ITS) used in the large-scale evaluation study had finished at the time of writing the previous deliverable for this work package (D6.2), the current deliverable, instead of describing new technical developments, discusses the evaluation study conducted using this system, with a focus on the output module in the context of storytelling. In addition, the results of two further studies related to output generation are presented.

2 Large-scale study and the role of output generation

In the period between February and June 2018 a large-scale study was conducted, where a total of 194 children followed all seven lessons with the L2TOR ITS. The study was preregistered [2], with the following hypotheses:

H1 The robot will be effective at teaching children L2 target words: children will learn words from a robot (H1a) and will remember them better (H1b) than children who participate in a no treatment (control) condition.

H2 Children will learn more words (H2a), and will remember them better (H2b) when learning from a robot than from only a tablet.

H3 Children will learn more words (H3a), and will remember them better (H3b) when learning from a robot that produces iconic gestures than from one that does not produce such gestures.

Further details regarding the design of this experiment, as well as its findings, are described in Deliverable 7.2. A paper documenting the study has been accepted for publication at the upcoming

International Conference on Human-Robot Interaction (HRI 2019) [3]. In the current deliverable, we focus on Hypothesis 3 because it relates directly to the non-verbal parts of the output module. In order to test this hypothesis, two experimental conditions were included in the study: one where the robot performed only deictic gestures to guide the learner's attention, and one where the robot additionally performed an iconic gesture every time it presented a target word in the second language (L2). Based on existing literature on the positive effects of (congruent) iconic gestures in human-human second language teaching (e.g., [4]), as well as our previous findings regarding the robot's use of iconic gestures to support learning [1], we expected children in the experimental condition with iconic gestures to perform better at learning second language vocabulary than children who did not get the support of a robot performing iconic gestures. However, the results of the study do not confirm this hypothesis, therefore we cannot conclude that the robot's use of iconic gestures contributed to children's learning of new English vocabulary.

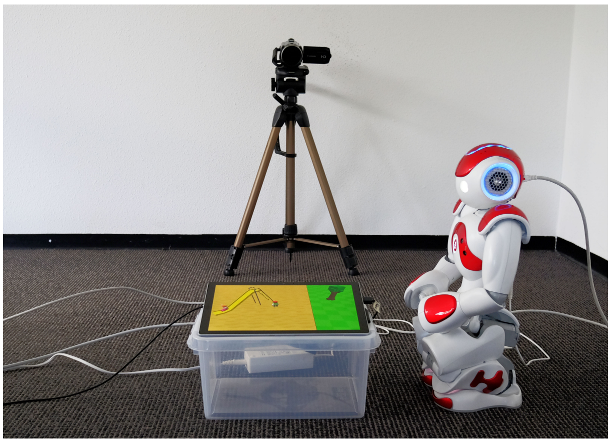


Figure 1: The set-up of the large-scale study (taken with permission from [3]).

Because these findings did not correspond with what existing literature and our previous research have indicated, we have set out to investigate which factors may have affected these results. Several potential causes are presented in the following sections.

2.1 Set-up of the experiment

First, the set-up of the environment and positioning of the robot are different from our previous experiments. In the current study, the robot is sitting down close to the child, at a 90 degree angle (Figure 1), while previously the robot was standing further away and opposite the child. This imposed limitations on the robot's ability to gesture, because only upper body motion could be used. Furthermore, the fact that the robot was not directly facing the child may have affected the way gestures were perceived, especially when the gesture relied on this perspective to obscure certain parts from view, for example in the case of finger counting (Figure 2). Anecdotally, we have noticed that children reproduced a gesture for *three* when in fact the robot was showing its gesture for *two*, an indication that trying to hide the thumb from view did not always work. What appears to be a *positive* result of this change in set-up, however, is that children tended to spontaneously re-enact the gestures that were performed by the robot more often in the recent study. This could be due to the gestures being relatively less complex, and because the robot and child are both sitting, so the child would not have to get up in order to perform the gesture. We are planning to investigate whether this subgroup of children that spontaneously re-enacted gestures performed better on the post-tests than children that only observed the gestures from the robot without copying them.

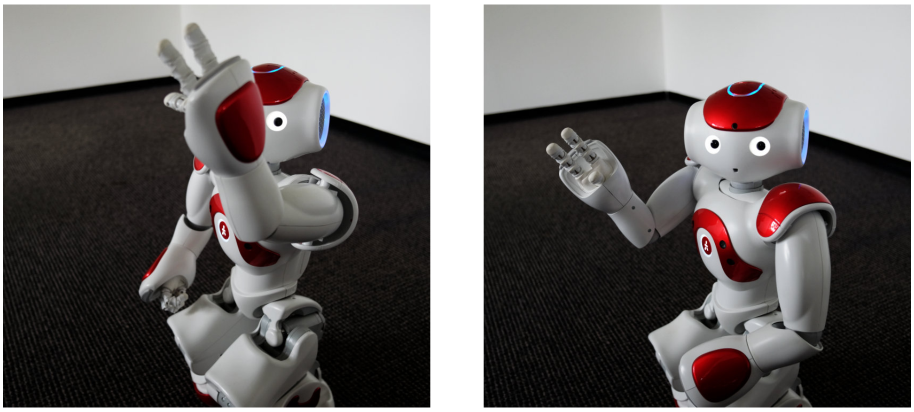


Figure 2: The robot performing gestures for *two* (left) and *three* (right).

2.2 Design of the gestures

Combined with the robot's restricted ability to gesture, the increased number of 34 target words made it more difficult to design gestures that are unique enough to avoid confusion. Furthermore, not all gestures may have been equally iconic, because the target words included relatively abstract concepts such as spatial relations. For example, it is easier to come up with iconic gestures for action verbs such as *running* or *flying*, or the animal names that were used previously, than for concepts such as *add*, *high*, or *behind*. To further investigate the clarity and iconicity of the gestures that were created, we are planning to conduct a perception study where participants will be asked to view a gesture performed by the robot, recorded from the same 90 degree angle while sitting as in the study (shown in Figure 2). All target words from the same lesson are then presented, six in total, from which the participant is asked to choose the correct word to which the gesture belongs. This will verify whether the gestures were at least unique enough within the lesson in which they were used.

One further assumption that was made when implementing the gestures, is that all children benefit equally from them. It is possible that some form of personalisation is needed, for example by only showing gestures when they relate to a word with which the child is currently struggling, as an additional support mechanism. For experimental consistency, the current implementation includes a gesture every time a target word is mentioned by the robot, which adds up to a large number of exposures to the same gesture and, as a result, adds substantial delays to the flow of the interaction between child and robot. Furthermore, there was no variation in the gestures performed for a particular concept – characteristics such as strategy used, speed, size, and complexity of the gestures were kept the same throughout the interaction. We have reflected upon these design decisions in Deliverable 6.4 and in a paper that was recently presented at a workshop on Natural Language Generation for Human-Robot Interaction, at the International Conference on Natural Language Generation (INLG2018), which was hosted in Tilburg [5].

2.3 Role of the tablet

A final circumstance that may have mitigated the contribution of the robot's gestures is that a large part of the interaction between the child and the robot actually took place within a virtual environment on the tablet. This move towards a tablet game was a conscious choice because it is still challenging to implement socially intelligent behaviour for a robot, in an unconstrained physical environment [6]. It is conceivable however that any issues that might occur while interacting with the tablet will have also affected the experience of the system as a whole. If too much of the child's attention had indeed been directed towards the tablet, for example due to issues with, or complexity of the interaction, this may have distracted from the actions performed by the robot (such as gestures). In order to investigate whether this was indeed the case, we conducted an evaluation of the usability and user experience of the ITS as a whole. To get a comprehensive overview, three different evaluation methods were combined: observations on video recordings of the large-scale study, expert reviews with students in the field of human-computer interaction, and a test session with older children (11–12 years old). The results of this evaluation show that the majority of the issues with the system can indeed be traced back to the interaction with the tablet, although issues relating to the speech and gestures of the robot were reported as well. We have written a paper, which is also appended to this deliverable, where we discuss the concrete issues that we have encountered, and give advice to include any external devices, such as the tablet game, in an iterative, user-centered design process.

3 Encouraging production of spatial concepts in an L2

As one of our continued small-scale studies we looked at the potential of using robots to assess a child's productive learning gains in a second language. Previous research has shown that receptive vocabulary tends to be bigger than productive vocabulary in first language (L1) [7, 8], and that L2 learners obtain lower scores on productive tests as compared to receptive tests [9]. This has been formalised into a hierarchy for word knowledge [10].

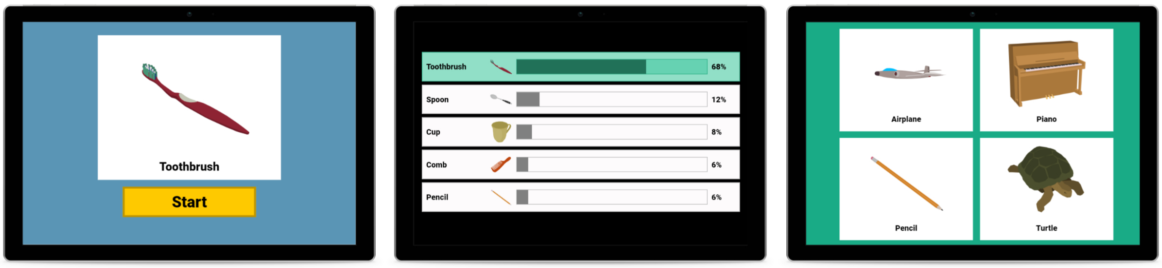
We wanted to take advantage of a robot's previous shown ability to reduce foreign language anxiety [11] to better assess a child's productive vocabulary. We designed a study using a robot which involved us teaching children some spatial French words (*sur*, *sous* and *devant*). Which we then tested them as part of a quiz game.

While we were unable to surpass human performance at assessing a child's productive vocabulary, we did find no significant difference between the two. There were potentially still improvements we could make to the implementation of the robot that could further improve on this performance. We also concluded that the process of testing vocabulary was long and repetitive, as such a robot may be suitable as a tool for teachers, to alleviate some of their time constraints.

4 Comprehensibility of recorded gestures

In Deliverable 6.2, we described a study with the goal of recording participants using a depth sensor (Kinect) while they were performing gestures. These gestures could then be mapped to the robot, enabling it to perform movements that it had previously ”learned” from humans. By using this learning by demonstration approach [12], the robot's gestures can be made to look more human-like and, if the set of recorded gestures is large enough to capture the variations in strategy, can be representative of how humans would approach the gesture production process. However, due to its physical limitations not all movements are expected to map equally well to the robot. Therefore, it is also important to get an indication of whether these recorded gestures still succeed in conveying the message for which they were originally intended.

In order to collect a comprehensive set of gestures, recorded in a naturalistic setting, we created an experiment that is modelled after the game of *charades*. Participants were asked to perform a gesture for a specific object that was shown to them on the tablet screen (Figure 3, left), after which the robot attempted to guess which object it was (Figure 3, middle). In the meantime, the gesture was also recorded and stored in the robot's memory. After guessing, it was the robot's turn to perform a gesture, which was done by automatically mapping one of the gestures recorded from a previous participant onto the robot, by converting the joint positions recorded by Kinect onto joint angles that the Nao robot accepts for performing its motions. The participant was then presented with four objects on the tablet screen (Figure 3, right), including the correct answer — the object for which a gesture was shown by the robot. If the participant guessed correctly, this is an indication that the particular gesture maps well to the robot's physical limitations, preserving its original meaning.



Toothbrush	68%
Spoon	12%
Cup	8%
Comb	6%
Pencil	6%

Figure 3: Left: the participant is shown an object to depict using upper body gesture. Middle: after guessing, the robot shows alternatives it was considering — the top ranked object is actually guessed. Right: after the robot performs a gesture, the participant is asked to guess.

The experiment ran fully autonomously, although researchers did have a button to mark the end of a participant's gesture in case the system did not manage to detect this automatically. Each session started with a practice round, where the robot and participant would each take one turn. During this introduction, the objects were always the same and robot would always guess correctly, because the goal of this round was to introduce the mechanics of the game. This first round was followed by five ”real” rounds, where the items were picked randomly out of a set of 35 objects, and the robot was actually guessing which object the participant was showing by finding the closest match in its set of existing recordings. The set of objects included musical instruments, animals, static objects (e.g., bridge, table), vehicles, and tools (e.g., cup, toothbrush). Figure 4 shows the set-up of the experiment.

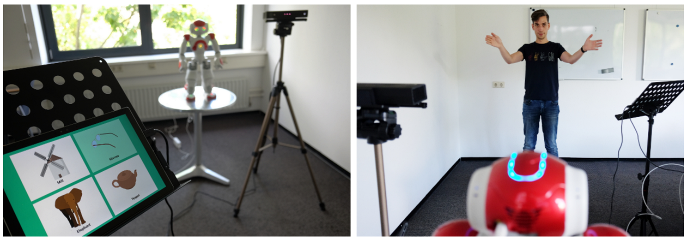


Figure 4: Left: the participant correctly guessed a gesture that the robot just performed. Right: the participant is performing a gesture (*ball*) for the robot to guess.

An identical version of the experiment was set up at the NEMO science museum in Amsterdam for fourteen days and at the Lowlands music festival in Biddinghuizen, the Netherlands, for three days. This resulted in approximately 3,500 recorded gestures from over 400 different participants. In between the two studies at different locations, the robot's memory of recorded gestures was reset in order to restart its learning process. A brief initial paper describing the experiment, which is also attached to this deliverable, has been accepted as a Late Breaking Report to the International Conference on Human-Robot Interaction (HRI 2019), with the aim to generate an interest in this line of research within the HRI community. We intend to follow up with a journal paper targeting a broader audience, accompanied by open access to the full dataset of recorded gestures, as well as the (modular) source code of the experiment. The dataset contains recordings in 3D of participants performing gestures for the 35 included objects, along with additional (pseudonymized) demographic data regarding the performer of each gesture. Figure 5 shows examples of different examples of the gesture for *guitar*.

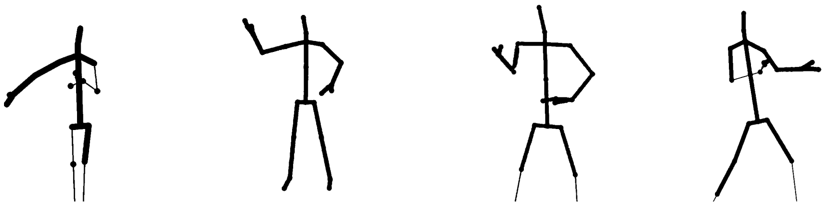


Figure 5: Different recordings of a gesture for *guitar*. The first two examples are performed by children, the last two by adults.

5 Conclusion and Future Directions

In this deliverable we have described the performance of the output module during the large-scale evaluation study. The storytelling component was represented throughout the seven lessons by creating a theme, including various scenes and narratives, for each of these lessons. Although the results of this study show that the system succeeds at teaching children new English vocabulary words, the robot's use of iconic gestures did not significantly improve the students' performance. In Section 2, several aspects of the design of this study are outlined which may have mitigated the positive effects that gestures could potentially bring. Future work includes measuring children's engagement with the robot and the (learning) tasks, to investigate whether there are differences between the experimental conditions with and without iconic gestures. Furthermore, we intend to investigate whether children that re-enacted the robot's gestures performed better than those that did not.

In addition, we present two further studies. The first study focused on encouraging the production of spatial concepts — which were also part of the large-scale study — in a second language, while trying to take advantage of the robot's ability to reduce foreign language anxiety. The second study was exploratory in nature, with the goal to generate a large, diverse dataset of recorded gestures. This dataset can be used for research into gestures in general, and can also be applied to the design of human-robot interactions. Both studies provide valuable input that can be used to improve future iterations of the Intelligent Tutoring System.

References

[1] Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. The effect of a robot's gestures and adaptive tutoring on children's acquisition of second language vocabularies. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 50–58. ACM, 2018.

[2] Rianne van den Berghe, Mirjam de Haas, Emiel Krahmer, Paul Leseman, Ora Oudgenoeg, Josje Verhagen, Paul Vogt, Bram Willemsen, and Jan de Wit. Aspredicted.org preregistration — L2TOR (#8181). http://web.archive.org/web/*/https://aspredicted.org/6k93k.pdf. Accessed: 11-12-2018.

[3] Paul Vogt, Rianne van den Berghe, Mirjam de Haas, Daniel Hernández García, Laura Hoffmann, Junko Kanero, Ezgi Mamus, Jean-Marc Montanier, Cansu Oranç, Ora Oudgenoeg-Paz, Fotios Papadopoulos, Thorsten Schodde, Josje Verhagen, Christopher D. Wallbridge, Bram Willemsen, Jan de Wit, Tony Belpaeme, Tilbe Göksun, Stefan Kopp, Emiel Krahmer, Aylin C. Küntay, Paul Leseman, and Amit K. Pandey. Second language tutoring using social robots: A large-scale study. To appear in *the 2019 ACM/IEEE International Conference on Human-Robot Interaction*, 2019.

[4] Spencer D Kelly, Tara McDevitt, and Megan Esch. Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and cognitive processes*, 24(2):313–334, 2009.

[5] Bram Willemsen, Jan de Wit, Emiel Krahmer, Mirjam de Haas, and Paul Vogt. Context-sensitive natural language generation for robot-assisted second language tutoring. In *Proceedings of the Workshop NLG4HRI at INLG 2018*. ACL, 2018.

[6] Guang-Zhong Yang, Jim Bellingham, Pierre E Dupont, Peer Fischer, Luciano Floridi, Robert Full, Neil Jacobstein, Vijay Kumar, Marcia McNutt, Robert Merrifield, et al. The grand challenges of science robotics. *Science Robotics*, 3(14):eaar7650, 2018.

[7] Batia Laufer and T. Sima Paribakht. The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, 48(3):365–391, 1998.

[8] Batia Laufer. The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, 19(2):255–271, 1998.

[9] Jan-Arjen Mondria and Boukje Wiersma. Receptive, productive, and receptive + productive l2 vocabulary learning: What difference does it make? In Paul Bogaards and Batia Laufer, editors, *Vocabulary in a Second Language: Selection, Acquisition and Testing*, pages 79–100. John Benjamins Publishers, 2004.

[10] Batia Laufer and Zahava Goldstein. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54(3):399–436, 2004.

[11] Minoo Alemi. General impacts of integrating advanced and modern technologies on teaching english as a foreign language. *International Journal on Integrating Technology in Education*, 5(1):13–26, 2016.

[12] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Why User Experience Research Matters for Human-Robot Interaction: The Case of Tablet Use in an Intelligent Tutoring System

Jan de Wit
*TiCC**
Tilburg University
Tilburg, the Netherlands
j.m.s.dewit@uvt.nl

Laura Pijpers
*TiCC**
Tilburg University
Tilburg, the Netherlands
laura.pijpers93@gmail.com

Rianne van den Berghe
Department of Special Education
Utrecht University
Utrecht, the Netherlands
m.a.j.vandenberghe@uu.nl

Emiel Krahmer
*TiCC**
Tilburg University
Tilburg, the Netherlands
e.j.krahmer@uvt.nl

Paul Vogt
*TiCC**
Tilburg University
Tilburg, the Netherlands
p.a.vogt@uvt.nl

***Abstract*—Many human-robot interaction systems involve a third component: a tablet, which can either be separate or integrated in the robot (as is the case in Softbank's Pepper robot). Such a tablet can be used, for instance, to present information to the human user or to gain control over the robot's complex surroundings, by introducing a virtual environment as a substitute for interactions that would normally happen in the physical world. While such a tablet can potentially have a big impact on the usability of the entire system and affect the interaction between human and robot, they are often not explicitly included when evaluating the user experience of human-robot interaction. This paper describes a case study where three evaluation methods were combined in order to get a comprehensive overview of the user experience of an Intelligent Tutoring System (ITS), consisting of a robot and a tablet. The results show several major usability issues with the virtual environment, which could have affected the experience of interacting with the robot. This underlines the importance of including not only the robot itself, but also any other interaction mediators in an iterative design process.**

***Index Terms*—Autonomous robots, User centered design, Design methodology, Human-robot interaction**

I. Introduction

A. Social Robots and the Need for Autonomy

Robots are increasingly being used for application domains in which they are expected to interact frequently with humans, and thus to exhibit socially intelligent behavior. Examples of such domains include personal assistance, education, and health care [1]. Socially intelligent behavior relates to aspects such as expressing and perceiving emotions, communicating with high-level dialogue, establishing and maintaining social relationships, using natural cues such as gaze and gestures, showing personality and character and displaying the ability to learn or develop social competencies [2]. Bartneck and Forlizzi have proposed the following definition of social robots [3]:

> "A social robot is 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."

If a robot is not at least partially responsible for its own socially intelligent behavior, we are not interacting with the robot itself but rather using the robot as a medium to enable human-human interaction with the person that is controlling it (i.e., telepresence). Therefore, a social robot will need to exhibit at least a certain degree of autonomous behavior. Based on a literature review, Beer et al. define robot autonomy as [4]:

> "The extent to which a robot can sense its environment, plan based on that environment, and act upon that environment with the intent of reaching some task-specific goal (either given to or created by the robot) without external control."

The combination of these definitions indicates that interacting with humans in their natural environment is inherently part of the task-specific goals an autonomous social robot aims to achieve. In order to reach these goals, on top of existing task-specific robot behaviors there is an additional need for the robot to be able to sense social cues, and to plan and act in a socially intelligent way. This will result in robots that can be deployed in a range of real world settings.

Social robots that are able to operate fully autonomously are beneficial for research in HRI as well. For instance, studies conducted with robots in the field — at home, school, health care facilities — will lead to results with higher ecological validity than studies that are being done in controlled lab

This research was funded by the EU H2020 L2TOR project (grant 688014).

* Tilburg center for Cognition and Communication

settings. Furthermore, there is an increasing need to investigate whether the effects we see with robots on the short-term persist over longer periods of time [5]. If the robot is able to perform autonomously, it will be possible to conduct long-term experiments in the wild without the substantial time investment of having a researcher control the robot's behavior, which in turn reduces bias and improves the replicability of human-robot interaction studies.

B. Challenges When Designing Autonomous Behavior

Although autonomous robots have already been deployed successfully in the field of industrial automation, introducing an advanced level of autonomy to social robots is challenging because the environment in which they operate is more dynamic and less constrained than that of industrial robots. The sensing abilities of social robots include the observation of not only the robot's often complex and unpredictable physical surroundings, but also the characteristics and behavior of humans that are present in this environment. Human social behavior is a complex phenomenon that is still under research, which adds to the challenge of developing a robot that is able to interact socially with humans [6]. It is unclear exactly which types of sensing, planning and acting functionalities are needed to facilitate social interactions. Moreover, some of the techniques that are currently being used do not perform well enough to be used autonomously, in a complex environment. For example, it is challenging to automatically keep track of arbitrary physical objects within an environment, without either augmenting the objects or restricting the environment [7]. Similar challenges occur regarding the sensing of auditory cues such as unconstrained speech in a noisy environment [8], especially with young interlocutors [9]. Human characteristics of social interactions, such as the level of engagement of the conversational partner, are multifaceted and therefore difficult for a robot to gauge [10]. Subsequent planning and acting steps are thus often based on abstract or incomplete information. The actions that the robot is expected to produce include physical interactions with the environment, as well as elaborate social behaviors outlined in I-A, such as maintaining a dialogue. Robots that are commonly deployed in social contexts today are limited when it comes to performing these actions: the robot's speech can contain imperfections and might lack emotion [11] and, due to its limited degrees of freedom, the robot is not able to gesture as fluently and with as much detail as humans [12]. Furthermore, social robots are likely to struggle when trying to manipulate their physical environment [13]. Finally, successful completion of the robot's task-specific goals can be difficult to measure, when these goals involve a change in knowledge state, attitude or behavior of a conversational partner.

To summarize: we are at a point in time where the various aspects of a robot's autonomy — sensing, planning and acting — are still challenging to implement when its tasks involve operating in a dynamic, social context. There are two ways in which we tend to cope with these on-going challenges. The first one is the use of Wizard of Oz techniques, where a researcher is fulfilling (parts of) the robot's sensing or planning abilities, essentially providing it with the input it needs in order to perform its actions. This is useful when trying to investigate how people would respond to a robot's (social) behavior, with the assumption that robots of the future would be able to sense and plan without the Wizard. Ideally, the Wizard only substitutes those functionalities that robots would realistically be able to do in the near future [14], for example by making decisions for the robot (planning), but only based on information that results from the robot's implemented sensing abilities. The potential pitfall of this method is that researcher bias could be introduced into the experiments, and this in turn may lead to results that are difficult to reproduce. As we will show later, it also affects the validity of several usability evaluation performance measures, such as time on task, when the Wizard has a role in the user's completion of the task.

The second option to achieve autonomous behavior in a complex environment is to control or constrain this environment, thus making it easier for the robot to sense and act within its surroundings. This can be done by moving part of the human-robot interaction into the virtual domain, for example by introducing a tablet device as a mediator. Objects within the virtual space can easily be tracked and manipulated programmatically by robots, as well as through a graphical user interface by humans, thereby allowing both parties to collaborate on the device in order to complete their tasks. By moving tasks into the virtual domain, it becomes easier to measure whether they were completed successfully, and to manage the flow of the interaction. However, what is often not critically evaluated is how the introduction of such a virtual environment may influence the overall experience of the human-robot interaction, and to what extent it diminishes the benefits of the robot's physical presence in a natural context.

C. User Experience and HRI

User experience is described by Hartson and Pyla as [15]:

> "The totality of the effect or effects felt by a user as a result of interaction with, and the usage context of, a system, device, or product, including the influence of usability, usefulness, and emotional impact during interaction and savoring memory after interaction."

The concept of user experience has recently been investigated in the context of HRI, where three major challenges were identified [16]. The first challenge is the need for an iterative design process, which is relatively difficult to achieve due to the high costs of rapid prototyping, variations in interactions when a robot performs autonomously, and the complexity of an engineering process that includes many hardware and software components. Second, there is a need to define user experience goals at the onset of a project, that focus specifically on the quality of the interaction between human and robot, rather than on specific behaviors or features of the robot. These goals can then be used as a guideline throughout the development and evaluation of the system. Finally, there should be an awareness of the different user experience evaluation methods that exist, and their potential applications within the field of HRI, as well as methods that have been created or adapted specifically for evaluating human-robot interactions such as the USUS framework [17].

These challenges, by extension, should also apply to the measures that are being taken in order to facilitate autonomous behavior, discussed in I-B. The use of Wizard of Oz or the introduction of a virtual environment will have an effect on the user experience, either by directly becoming part of the system that is being 'experienced' (if the robot and virtual environment are tightly integrated and co-dependent, for example), or by being present within the context of use and thereby influencing the way human and robot interact. If such measures are being used, they should also be involved in an iterative design process, included when setting user experience goals, and subject to user experience evaluation methods.

D. Case Study: Longitudinal Second Language Tutoring

As part of the L2TOR project, an Intelligent Tutoring System (ITS) was developed for second language tutoring. The system, consisting of a SoftBank Robotics NAO robot and a tablet, was deployed at several primary schools to provide one-to-one tutoring. Figure 1 shows an overview of the setup of this tutoring interaction.

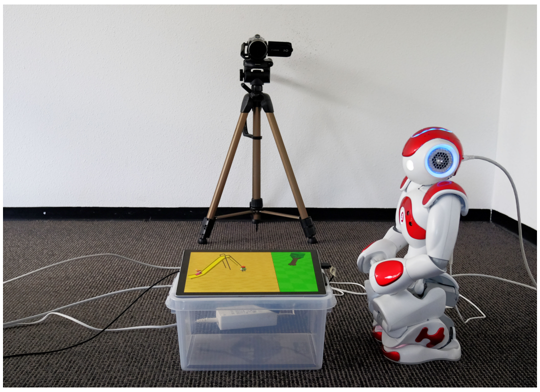


Fig. 1. The setup of the Intelligent Tutoring System (published with permission from [18]).

By interacting with the system, children of five to six years old were taught second language vocabulary, on average six words per lesson over the course of six lessons, followed by a final seventh lesson to repeat all terms from the previous six lessons. The words were selected based on literature regarding word learning in young children, and were distributed among lessons to gradually go up in difficulty. The robot guided the child in the learning process, taking on the role of a knowledgeable peer [19]. During the course of each lesson, children engaged in a variety of tasks that were designed to support learning, most of which involved interacting with the tablet device:

- Touching an object on the screen;
- Moving an object to a desired location on the screen;
- Moving an object to collide with another object;
- Repeating a word or sentence after the robot;
- Performing a physical act (acting out a certain activity, showing left or right hand).

The total number of interactions, as well as the distribution of interaction types among lessons, is shown in Figure 2. This provides an indication of the difficulty of the lessons from an interaction design perspective. In lessons 4–6 the complexity of repeating after the robot became more difficult, as repetition moved from single words to short sentences. It is worth noting that the seventh (recap) lesson was vastly different from the other six in several key aspects: rather than using a 3D engine, this lesson used a 2D environment. In this scenario, children were creating a photo album together with the robot, where the background of each page was a screenshot of one of the previous lessons and the child was asked to put stickers on the page. Therefore, moving objects was in this case also done in two dimensions and there were no collisions between stickers.

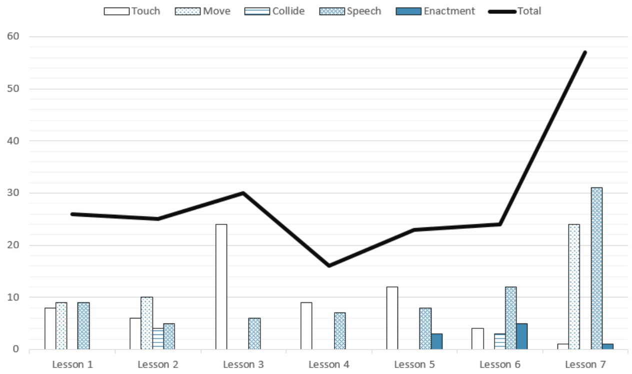


Fig. 2. The number of interactions per lesson, split by interaction type.

Children taking part in the study were assigned to one of four conditions: a control condition without exposure to educational content, a tablet condition where the robot was not present, a condition where the robot performed deictic gestures (to point at, and pretend to manipulate, objects on the tablet) only, and a condition where the robot performed deictic gestures and iconic gestures (related to the 34 target words). Further details regarding the study and its main results are described in [18].

The robot functioned fully autonomously, with the exception of validating tasks where speech recognition was required. In that case, a Wizard of Oz approach was used, which was limited to voice activity detection rather than full speech recognition. In other words, the researcher pressed a button when the child said *something*, and then the robot provided neutral feedback to praise the effort of the child. It was a conscious choice to only make the Wizard perform an action (voice activity detection) that is realistically feasible for robots to do in the near future [14], rather than full speech recognition. However, a number of design decisions were made to mitigate several of the challenges of autonomous behavior, outlined in I-B. For example, to avoid a negative effect of imperfections in the robot's pronunciation, each initial exposure to a new word was done by playing back a sample of a recorded native speaker from the tablet when first introducing the target word. Furthermore, most of the interactions took place in a virtual environment on the tablet, which was developed specifically for this study. This removes much of the unpredictability and complexity of a real, physical environment, thus giving the robot the ability to easily sense and act within this context. Interactions with this environment were also carefully scripted and choreographed, for example by locking objects from being manipulated until a task was presented that required interacting with them. The only sensing that the robot performed in the physical world was to track the child's face in order to establish eye contact. Tasks that required the child to perform an activity in the real world were not validated by the robot; it would simply wait for a certain period of time before continuing with the lesson. There was also no personalization of the interaction based on the child's (non-verbal) input, mental state, or estimated knowledge level. An example of a virtual environment that was used during the lessons is shown in Figure 3.

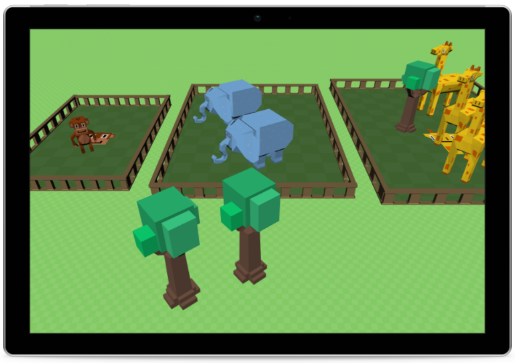


Fig. 3. The virtual environment used in lesson one of the experiment.

E. *In this Paper*

We present a usability and user experience evaluation of an Intelligent Tutoring System (ITS) with an autonomous humanoid robot, applied in a longitudinal tutoring interaction with pre-school children. This work highlights the importance of conducting such evaluations when designing a Human-Robot Interaction (HRI), particularly when (custom-made) external peripherals that affect this interaction, such as tablet games, are involved. Based on our concrete usability findings, we propose general design guidelines, as well as suggestions to better shape the design process for HRI applications.

II. Methodology

In order to get a comprehensive overview of the state of the usability and user experience of the ITS described in I-D, we used a triangulation approach [20] to combine the outcomes of three different evaluation methods. First, observations were conducted on a set of recordings of child-robot interactions from the study described in I-D, and metrics were extracted from the log files belonging to these interactions. Second, design experts were asked to evaluate the system based on a set of heuristics. Finally, the ITS was evaluated with older children by means of a semi-structured interview. We will discuss each of the three evaluation methods separately in the following subsections.

The combined evaluation resulted in quantitative measurements of children's performance, qualitative feedback on the user experience and a list of usability issues. To consolidate the reported severity ratings of each usability issue from multiple methods, a Damage Index [21] was calculated as follows:

$$DI = \frac{\bar{s} * n}{y * N} \tag{1}$$

In this formula, $\bar{s}$ is the mean severity rating of the issue, over all methods in which the issue was uncovered, n the number of evaluations (methods) in which the issue was encountered, y is the upper bound of the severity scale (in this case, 4) and N is the number of evaluations conducted (3 in this study). Because the severity ratings used in this evaluation range from 1 (worst) to 4 (best) and the Damage Index expects a higher number to be worse, the ratings were inverted before calculating the Damage Index. The measure will result in higher ratings for those issues with high average severity ratings, and that have come up during multiple (different) evaluation methods. This last factor could be seen as a limitation, as difficult issues to identify (e.g., those that are context-sensitive or timing-related) may receive a lower Damage Index even if their severity is high. However, one could argue that these issues are relatively rare and should therefore indeed get less priority than those that affect a larger group of people, even if they have a lower severity rating.

A. *Observations and Log Files*

The observations were done by one of the researchers. Due to time constraints, a random selection was made to include video recordings of twenty children from each of the three experimental conditions (tablet only, tablet + robot with deictic gestures, tablet + robot with deictic and iconic gestures). Two lessons were included, lessons one and six, to also take into account learnability of the system. This resulted in a total of 120 videos (60 children, two lessons) of approximately twenty minutes each. Log files of the same children, for the same two lessons, were used to derive objective performance measures.

Each usability issue that occurred was noted down, as well as any utterances from the child towards either the experimenter or the robot. After all observations were completed, the usability issues were categorized and a severity rating was assigned. Four categories of severity proposed by Dumas and Redish [22] were used, where 1) prevents task completion, 2) creates significant delays and frustration, 3) has a minor effect on usability, and 4) are subtle suggestions for future enhancements. From the log files, the time on task, task success and number of errors were extracted by means of an automated script.

B. Heuristic Evaluation

Three participants with knowledge about usability and user experience, one male and two female, with an average age of 24 years (*SD* = 1 year) were recruited through purposive sampling. They were all master's students enrolled in a program related to Human-Computer Interaction. Because the emphasis of the study was on measuring the experience of using the entire ITS, only the two conditions including the robot were evaluated and the same two lessons from the observations were examined. Two experts played lesson one with the robot performing both deictic and iconic gestures, and lesson six with the robot using only deictic gestures. For the third expert, the two conditions were reversed (lesson one with only deictic gestures, lesson six with deictic and iconic gestures).

Although the interaction could be paused and resumed at any time, there was no option to freely navigate through the lesson content. Therefore, participants in the heuristic evaluation were provided with printed screenshots of the 3D tablet environments used in the lessons. Print-outs of the Heuristic Evaluation Child E-learning applications (HECE) [23] were given as a guideline. The HECE is an extension of the traditional user interface principles by Nielsen [24] with additional heuristics related to usability specifically for children and e-learning. After writing down a list of issues and relating them to these heuristic principles, participants were asked to assign a severity rating, from the same scale as the one used during the observations [22], to each issue.

C. Usability Study and Interview

For the usability study and interview, convenience sampling was used to recruit ten children, four male and six female, with an average age of eleven years and eight months (*SD* = 6 months). We chose to conduct this part of the evaluation with children that were older than the intended age of the ITS, because these children have reached a developmental stage where they can consider ideas, multiple solutions and hypothetical situations [25], while still being in the primary school context. This enables them to imagine what the experience would be like for their younger peers.

During the usability study, which took place at the participants' primary school, only one lesson (either lesson one or six) was shown to each participant, in the robot with deictic and iconic gestures condition. Due to time constraints, there were four pairs of two children conducting the study together, and the last two children participated individually so that lesson one and six would each be rated five times. The children were asked to go through the lesson while thinking aloud. Several pre-determined questions about specific parts of the system (e.g., regarding the design of the on-screen characters and environments) were posed during the interaction, and afterwards the participants were asked to give their general opinion about the ITS: what they liked and disliked, whether they thought it was suitable to teach children aged 5–6 a second language, if the tablet game would also be fun without the robot present and what they would do differently given the chance to redesign the system. The result of this study was again a list of usability issues, combined with qualitative feedback regarding older children's attitudes towards the system.

III. Results

A. Observations and Log Files

The observations resulted in a list of usability issues for lessons one and six, along with the number of times in total, and for how many individual children, each issue occurred. The issues were then grouped into the following five categories: output from the agent (e.g., providing negative feedback when a task was actually completed successfully), design of the tablet game (e.g., the screen would turn black to guide the child's attention, but this was never properly explained), actions performed by the system (e.g., agent moves the wrong object on the screen), issues with the content (e.g., tasks not clearly introduced), and other issues (e.g., error message unrelated to the system appeared on the tablet). Table I shows an overview of the number of issues in each category, per lesson.

TABLE I
Number of issues discovered per category in both lessons

Category	Lesson one	Lesson six
Output from agent	24	4
Design of the tablet game	25	15
Actions performed by the system	3	0
Issues with the content	2	5
Other issues	1	1

There was no significant difference in the number of issues that occurred between experimental conditions, in either lesson. Two researchers assigned a severity rating to each issue. A total of 44 out of the 80 issues (55%) had a severity rating of 1 (prevents task completion) or 2 (significant delays and frustration). All of these were related to tasks where the child had to move an object to a new location, or to collide it with another object: either the robot was unaware of an object's current location (resulting in incorrect feedback from the robot), the robot manipulated the wrong object when providing help, objects were still locked when the child tried to move them, or the movement itself did not go smoothly. This last point was sometimes caused by children struggling to move objects on the screen in one go, occasionally losing the object half-way — this also resulted in negative feedback.

Eleven children from the observed videos (out of 60) made a total of 26 remarks during lesson one. Two children addressed the researcher, while the other nine children directed their remarks to the agent (tablet or robot, depending on experimental condition). Thirteen remarks were related to tasks and feedback of the system, such as *"That does not work"*. Other comments concerned the setting of the lesson, whether this was the only game there is, or that the lesson was taking quite long. For lesson six, thirteen remarks were made by nine children. Six children spoke to the researcher, while three addressed the agent. This time, six remarks were related to the interaction

(*"All right Robin, this is going to be hard"*); other topics included the real world environment where the experiment took place, and the camera used to record the experiment.

Performance measures were extracted from the log files collected by the system, for the same set of sessions included in the observations. The measures include time on task, measured in seconds from the moment the task was introduced by the agent until the moment the task was successfully completed, number of errors made (which influences time on task), and task success. The robot would always complete the task for the child after two erroneous attempts, or prolonged inactivity from the child — in this case the task was measured as being unsuccessful. Tasks where the child had to repeat or enact are omitted, because they were controlled by a Wizard of Oz or had a fixed duration, respectively. Both were also not evaluated for success, because the system could not observe correct task completion. Lesson one only contains tasks to move objects to a certain location and no collisions, while lesson six only has object collisions without tasks to move objects. It should also be noted that, because objects were colliding all the time with background scenery, errors were not registered for these collision tasks. The results of this analysis are shown in Table II, where the numbers of errors have been divided by the number of interactions of that type that were present in the lesson, to allow a fair comparison between lessons one and six.

TABLE II
PERFORMANCE MEASURES EXTRACTED FROM LOG FILES

	Lesson one	**Lesson six**
Avg time on task — touch	22.57 sec	7.70 sec
Avg time on task — move	28.99 sec	N/A
Avg time on task — collide	N/A	9.64 sec
Errors per interaction — touch	5	2.75
Errors per interaction — move	50.89	N/A
Errors per interaction — collide	N/A	N/A
Task success — touch	97.9%	98.4%
Task success — move	68.3%	N/A
Task success — collide	N/A	96.7%

B. Heuristic Evaluation

Participants in the heuristic evaluation were invited to create a list of usability issues they encountered, to link each issue to one of the heuristic guidelines that were provided and to assign a severity rating. A total of 25 issues came up, of which eighteen were not specific to a lesson, two concerned lesson one and five were related to lesson six.

Eleven issues (44%) received the highest two severity ratings. These issues were related to tasks that could not be properly carried out on the tablet (due to bugs), a lack of feedback for tasks where the child has to enact something (e.g., raising their right hand), unclarity in the robot's pronunciation of certain words, the imposed, slow pace of the interaction, the design choices regarding the tablet game (2D was suggested over 3D) and the lack of introduction of certain game mechanics (the screen turning black to guide the child's attention). Without having seen Figure 2, the experts noted that according to them lesson six contained less interactive elements than lesson one.

The evaluators were also asked to indicate any positive points about the system. They noted the positive and motivating attitude of the robot, realistic movements within the game as well as from the robot, a friendly appearance of the system and good use of colors. The robot tracking the child's face to establish eye contact was also highlighted as a positive feature.

C. Usability Study and Interview

During the usability study, a total of ten children gave feedback on the system while having a chance to go through one of the lessons. All participants were generally positive about the tutoring system, noting that this way of teaching is the future, it was nice to learn English with a robot and that it would be a suitable tool for teaching our intended age group of 5–6 years old. They agreed that the robot was the best part of the system and that the game would be less attractive without it. Specific elements that the children liked include the playground environment (lesson six), the visual design of the game, verbal and non-verbal feedback (i.e., the colors of the robot's eyes changing with positive feedback) and the interactions. They thought the game looked nice and structured and the objects were blocky but good.

Issues that the children encountered while interacting with the system and suggestions that they made were also noted down, and a severity rating was assigned to these points by two of the researchers. From a total number of forty issues, five were assigned a severity rating of 1 or 2 (12.5%). These were related to problems when moving objects, difficulty understanding the robot's speech and confusion about what kind of action was expected of the user. The other comments overlap to a large extent with the findings from the heuristic evaluation (e.g., unclarity of gestures, and overall pacing of the interaction).

D. Combined Results

The lists of usability issues resulting from the three different evaluation methods were combined, where overlapping issues were merged. This resulted in a total of 35 unique issues. To get an overview of the severity of each issue, taking into account the number of evaluation methods in which it was reported, and the severity that was assigned in these methods, a Damage Index was calculated. Table III shows the number of issues belonging to different Damage Index ranges.

TABLE III
DAMAGE INDEX FOR THE REPORTED USABILITY ISSUES

	0–0.1	**0.11-0.2**	**0.21-0.3**	**0.31-0.4**	**0.41+**
Number of issues	5	14	3	7	6

The top thirteen issues with a Damage Index of at least 0.31 were related to problems with dragging objects on the tablet, tasks not being clear to the user, the slow and fixed pacing of the interaction, limited control of the user over the system, unnatural and unclear speech from the agent, interaction mechanics not being properly introduced, ambiguous words or gestures, lack of feedback from the agent, and objects on the screen being locked while the agent is talking. The full results are made available online[1].

IV. Discussion

A. The Evaluation Study

The results of the evaluation show that the use of a virtual environment as a mediator for human-robot interactions can greatly affect the overall user experience. Most issues reported were either directly related to the interactions with the tablet (such as issues with moving objects on the screen), or listed as issues with the robot although they can actually be traced back to the tablet, because the robot was provided with incorrect information regarding the state of the virtual environment. This in turn resulted in the robot performing incorrect actions based on erroneous input, such as saying the wrong things or giving negative feedback when the task was in fact completed successfully.

Figure 2 shows that there was no gradual increase in difficulty of the tasks that the child had to perform during the lessons. In fact, moving objects on the screen was found to be the most challenging interaction, resulting in several reported issues, and this interaction type appears exclusively in the first two lessons. In later lessons, these tasks were implemented with variations where objects had to collide, or only had to be touched once rather than moved. However, the use of these different mechanics to achieve similar goals resulted in a reduced consistency of interactions throughout the lesson series, and prevented us from being able to evaluate all of the objective performance measures (listed in Table II). Furthermore, although the robot and its capabilities were introduced prior to the first lesson in a group setting, the mechanics of the tablet game were not included in this introduction. Future improvements to the system should focus on improving the interaction design, and a proper introduction of the workings of not only the robot but also the virtual environment in which it operates.

B. Evaluation Method

The present study shows the added value of combining three different approaches to evaluating the experience of interacting with a social robot, in order to get a comprehensive overview of issues with the system as a whole. Out of the 35 issues in the combined list presented in III-D, 22 were identified in only one of three evaluation methods. This is an indication that a substantial amount of issues would not have been identified if we had only done an evaluation with design experts, for example. However, in this case the evaluations were conducted *after* the system was already used in a large-scale experiment [18]. It is now impossible to tell whether, and to what extent, the findings from that experiment were influenced by the issues encountered when evaluating the system, without running a similar experiment after resolving these issues. In future work, we would therefore start evaluating earlier in, and more frequently throughout the design process. In the current study, the methods and measures that were used are taken from the broad field of human-computer interaction. Although these support our need to evaluate the tablet game and the ITS as a whole, to get a more comprehensive overview we would consider to combine the current methods used with those that are more specific to HRI, such as the USUS [17].

C. Autonomy, Mediators and User Experience

As we work towards creating social robots that are capable of operating fully autonomously, in a complex and dynamic environment, we currently still resort to Wizard of Oz approaches and ways to exercise control over the environment in order to deal with technical limitations and the intricacies of social interactions. Although the robot presented in our case study was able to display *some* socially intelligent behavior, a number of its sensing abilities were not implemented, or supported by a Wizard of Oz, and most of its actions were scripted rather than tailored to the situation. More importantly, although the robot did perform gestures within its physical environment, to a large extent the interactions with the learner took place in a virtual scene. Although we do believe that this is the way to move forward with HRI research, in extreme cases this could lead to a shift of attention towards these secondary objects and any usability issues they might contain, rather than the robot itself. Conscious design choices should be made, for example when delegating interactive elements to a virtual environment, and these should be well documented (e.g., as part of user experience goals) to preserve the added value that the robot can bring to the table.

V. Conclusion

This paper describes a case study in which a usability and user experience evaluation of an Intelligent Tutoring System was conducted. The study underlines the importance of evaluating the overall experience of a human-robot interaction, including any mediating devices that are introduced to gain control over the robot's environment in order to increase the robot's level of social autonomy. Furthermore, we urge researchers to allocate resources to the design and development of such interaction mediators, and to report exactly to which degree their robot is able to behave autonomously, as well as any concessions or work-arounds that might be in place. This would ensure that the effects of any mediators on experimental findings are minimized, while at the same time providing the HRI community with enough information to be able to reproduce these findings.

References

[1] K. Dautenhahn, "Socially intelligent robots: dimensions of human–robot interaction," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 362, no. 1480, pp. 679–704, 2007.

[2] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and autonomous systems*, vol. 42, no. 3-4, pp. 143–166, 2003.

[1] https://bit.ly/2E9UTK4

[3] C. Bartneck and J. Forlizzi, "A design-centred framework for social human-robot interaction," in *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication, RO-MAN*. IEEE, 2004, pp. 591–594.

[4] J. M. Beer, A. D. Fisk, and W. A. Rogers, "Toward a framework for levels of robot autonomy in human-robot interaction," *Journal of Human-Robot Interaction*, vol. 3, no. 2, pp. 74–99, 2014.

[5] P. Baxter, J. Kennedy, E. Senft, S. Lemaignan, and T. Belpaeme, "From characterising three years of HRI to methodology and reporting recommendations," in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 2016, pp. 391–398.

[6] G.-Z. Yang, J. Bellingham, P. E. Dupont, P. Fischer, L. Floridi, R. Full, N. Jacobstein, V. Kumar, M. McNutt, R. Merrifield *et al.*, "The grand challenges of science robotics," *Science Robotics*, vol. 3, no. 14, p. eaar7650, 2018.

[7] C. D. Wallbridge, S. Lemaignan, and T. Belpaeme, "Qualitative review of object recognition techniques for tabletop manipulation," in *Proceedings of the 5th International Conference on Human Agent Interaction*. ACM, 2017, pp. 359–363.

[8] R. K. Moore, "Spoken language processing: Piecing together the puzzle," *Speech communication*, vol. 49, no. 5, pp. 418–435, 2007.

[9] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Child speech recognition in human-robot interaction: evaluations and recommendations," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2017, pp. 82–90.

[10] T. Schodde, L. Hoffmann, and S. Kopp, "How to manage affective state in child-robot tutoring interactions?" in *Proceedings of the International Conference on Companion Technology (ICCT)*. IEEE, 2017, pp. 1–6.

[11] J. Barnes, E. Richie, Q. Lin, M. Jeon, and C. H. Park, "Emotive voice acceptance in human-robot interaction," in *Proceedings of the 24th International Conference on Auditory Display*, 2018.

[12] J. de Wit, T. Schodde, B. Willemsen, K. Bergmann, M. de Haas, S. Kopp, E. Krahmer, and P. Vogt, "The effect of a robot's gestures and adaptive tutoring on children's acquisition of second language vocabularies," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2018, pp. 50–58.

[13] C. C. Kemp, A. Edsinger, and E. Torres-Jara, "Challenges for robot manipulation in human environments [grand challenges of robotics]," *IEEE Robotics & Automation Magazine*, vol. 14, no. 1, pp. 20–29, 2007.

[14] N. Martelaro, "Wizard-of-oz interfaces as a step towards autonomous HRI," in *2016 AAAI Spring Symposium Series*, 2016.

[15] R. Hartson and P. S. Pyla, *The UX Book: Process and guidelines for ensuring a quality user experience*. Elsevier, 2012.

[16] J. Lindblom and R. Andreasson, "Current challenges for ux evaluation of human-robot interaction," in *Advances in ergonomics of manufacturing: Managing the enterprise of the future*. Springer, 2016, pp. 267–277.

[17] A. Weiss, R. Bernhaupt, M. Lankes, and M. Tscheligi, "The usus evaluation framework for human-robot interaction," in *AISB2009: proceedings of the symposium on new frontiers in human-robot interaction*, vol. 4, 2009, pp. 11–26.

[18] P. Vogt, R. van den Berghe, M. de Haas, L. Hoffmann, J. Kanero, E. Mamus, J.-M. Montanier, C. Oranç, O. Oudgenoeg-Paz, F. Papadopoulos, T. Schodde, J. Verhagen, C. D. Wallbridge, B. Willemsen, J. de Wit, T. Belpaeme, K. Bergmann, T. Göksun, S. Kopp, E. Krahmer, A. C. Küntay, P. Leseman, and A. K. Pandey, "Second language tutoring using social robots: A large-scale study," to appear in the 2019 ACM/IEEE International Conference on Human-Robot Interaction, 2019.

[19] C. Zaga, M. Lohse, K. P. Truong, and V. Evers, "The effect of a robots social character on childrens task engagement: Peer versus tutor," in *International Conference on Social Robotics*. Springer, 2015, pp. 704–713.

[20] C. E. Wilson, "Triangulation: the explicit use of multiple methods, measures, and approaches for determining core issues in product development," *Interactions*, vol. 13, no. 6, pp. 46–ff, 2006.

[21] G. Sim and J. C. Read, "The damage index: an aggregation tool for usability problem prioritisation," in *Proceedings of the 24th BCS Interaction Specialist Group Conference*. British Computer Society, 2010, pp. 54–61.

[22] J. S. Dumas and J. Redish, *A practical guide to usability testing*. Intellect books, 1999.

[23] A. Alsumait and A. Al-Osaimi, "Usability heuristics evaluation for child e-learning applications," in *Proceedings of the 11th international conference on information integration and web-based applications & services*. ACM, 2009, pp. 425–430.

[24] J. Nielsen, "Enhancing the explanatory power of usability heuristics," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 1994, pp. 152–158.

[25] P. Markopoulos, J. C. Read, S. MacFarlane, and J. Hoysniemi, *Evaluating children's interactive products: principles and practices for interaction designers*. Elsevier, 2008.

Using a Robot Peer to Encourage the Production of Spatial Concepts in a Second Language

Christopher D. Wallbridge
University of Plymouth
Plymouth, UK
christopher.wallbridge@plymouth.ac.uk

Rianne van den Berghe
Utrecht University
Utrecht, Netherlands
m.a.j.vandenberghe@uu.nl

Daniel Hernández García
University of Plymouth
Plymouth, UK
daniel.hernandez@plymouth.ac.uk

Junko Kanero
Koç University
Istanbul, Turkey
jkanero@ku.edu.tr

Séverin Lemaignan
Bristol Robotics Laboratory
Bristol, UK
severin.lemaignan@brl.ac.uk

Charlotte Edmunds
University of Plymouth
Plymouth, UK
charlotte.edmunds@plymouth.ac.uk

Tony Belpaeme
Ghent University/University of Plymouth
Ghent, Belgium
tony.belpaeme@ugent.be

ABSTRACT

We conducted a study with 25 children to investigate the effectiveness of a robot measuring and encouraging production of spatial concepts in a second language compared to a human experimenter. Productive vocabulary is often not measured in second language learning, due to the difficulty of both learning and assessing productive learning gains. We hypothesized that a robot peer may help assessing productive vocabulary. Previous studies on foreign language learning have found that robots can help to reduce language anxiety, leading to improved results. In our study we found that a robot is able to reach a similar performance to the experimenter in getting children to produce, despite the person's advantages in social ability, and discuss the extent to which a robot may be suitable for this task.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; • **Social and professional topics** → **Assistive technologies**; • **Computing methodologies** → Natural language processing; Cognitive robotics;

KEYWORDS

Robot Assisted Language Learning; Assessment; Second Language Learning

HAI '18, December 15–18, 2018, Southampton, United Kingdom

ACM ISBN 978-1-4503-5953-5/18/12…$15.00
https://doi.org/10.1145/3284432.3284433

ACM Reference Format:
Christopher D. Wallbridge, Rianne van den Berghe, Daniel Hernández García, Junko Kanero, Séverin Lemaignan, Charlotte Edmunds, and Tony Belpaeme. 2018. Using a Robot Peer to Encourage the Production of Spatial Concepts in a Second Language. In *6th International Conference on Human-Agent Interaction (HAI '18), December 15–18, 2018, Southampton, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3284432.3284433

1 INTRODUCTION

Learning the language of a new home region is vital for migrant children. It is beneficial for them to integrate with their peers, and necessary to prevent them from falling behind in school. Children need the opportunity to practice their language skills, but it may be difficult if no one at home is able to speak the language of the host region. Finding qualified teachers or tutors that know both the new language and the language of children's old homeland can also be challenging. With robots we may be able to support children's language learning needs.

When learning a second language (L2), it is difficult to master vocabulary both receptively and productively. L2 learners may find themselves capable of understanding the L2, while still struggling to produce L2 words. Indeed, previous research has shown that receptive vocabulary tends to be bigger than productive vocabulary in first language (L1) [8, 11], and that L2 learners obtain lower scores on productive tests as compared to receptive tests [14]. Thus, people are able to recognize more words than they can produce, both in their L1 and L2. This has been formalised into a hierarchy for word knowledge by Laufer et al. [9], based on knowing the words passively or actively and in being able to recognize them or recall them. The hierarchy is as follows, from easiest to most difficult: passive recognition → active recognition → passive recall → active recall. These are defined as follows:

- *Passive recognition* - The student is able to select the L1 word from a choice of words when provided the word in L2.

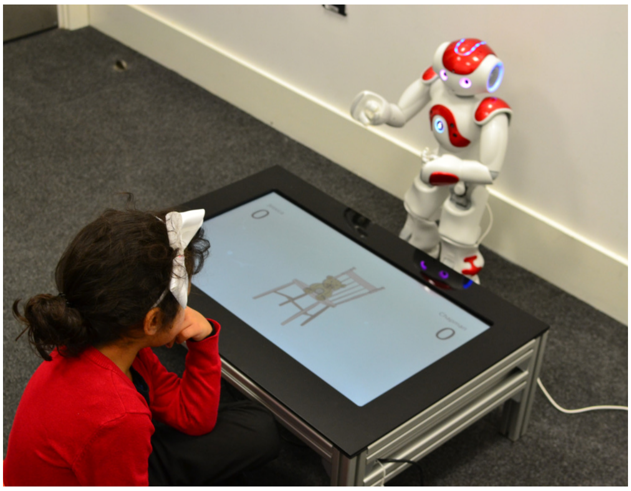


Figure 1: A child interacting with the robot in our study. The agent – in this case a robot – stands opposite from the child. An interactive table displays an image of a teddy bear and a chair. The child must use a word from a second language to describe the position of the bear in relation to the chair.

- *Active recognition* - The student is able to select the L2 word from a choice of words when provided the word in L1.
- *Passive recall* - The student is able to give the meaning of a word in L1 when provided the word in L2.
- *Active recall* - The student is able to give the L2 word when provided the word in L1.

This poses a challenge for L2 vocabulary interventions in which the trainer wants to assess the trainee's learning gains: L2 learners have difficulty learning the words productively (i.e. learning to produce foreign words), and will struggle to actively recall newly learned L2 words. There are several tests to assess an L2 learner's productive vocabulary, including assessments in which the participant has to describe pictures (e.g., the Expressive Vocabulary Test [18], the Expressive One-Word Picture Vocabulary Test [5], or the Clinical Evaluation of Language Fundamentals Test [17]), writing tests in which the learner has to fill in the blank (e.g., the Productive Vocabulary Levels Test [10]), or, for very young children, parental or teacher reports [4].

In many situations, it may not be possible to use one of these tests. For example, when the words learned concern abstract concepts, which cannot be easily depicted, it is not possible to use a picture test. If the learner is illiterate, one cannot use a writing test. Parents or teachers may struggle to report the child's L2 if they do not speak that language themselves. To further complicate the issue, producing L2 words may be intimidating for L2 learners. Even if the learner is able to produce the word, they may not produce it due to anxiety of pronouncing the word incorrectly [13].

A social robot may help overcome some of the issues described above in assessing L2 learner's vocabulary. While not being able to solve by itself the issue of vocabulary being more difficult to learn productively than receptively, a social robot may help in innovating novel ways to assess L2 vocabulary, or in reducing L2 anxiety in L2 vocabulary test settings. A robot may be less intimidating than an adult assessor, especially for young children, encouraging more speech production. This study evaluates whether school children may produce more L2 words in a productive L2 vocabulary test when playing with a social robot than with an adult. Below, we discuss relevant robot-assisted language learning (RALL) studies before detailing our study.

2 PREVIOUS WORK

RALL has been found to be effective in reducing foreign language anxiety (FLA), and teaching robots are able to improve oral skills of young students learning English as a foreign language [1]. Alemi et al. [2] performed a study using a robot teaching assistant. In the study, Persian-speaking students in Iran were taught English. A survey of the students showed that those who learned from the robot were significantly less anxious compared to the control group that did not have the robot. While a number of factors were thought to contribute to this reduction in anxiety, the authors claimed a major reason to be intentional mistakes the robot made. The mistakes not only gave the students a chance to correct the robot, but also made them less afraid of making errors of their own.

When looking at speaking skills, the focus can not just be on vocabulary gains, but pronunciation as well. Lee et al. [12] conducted a series of lessons to help Korean children from grades 3 to 5 (roughly 8 to 10 years old) learn English. In South Korea children start learning English from grade 3. As part of a lesson series they were given a pronunciation training with a robot, that used a lexicon that included often confused phonemes, so that the robot could correct the child's pronunciation. It was reported that the children's speaking skills improved significantly with a large effect size when measured by a teacher. As well as the improvement in speaking skills all three affective factors – interest, confidence and motivation – all improved significantly.

Instances of robots acting as care-receivers also occur in RALL. In a study by Tanaka and Matsuzoe [16], Japanese children were given the role of teaching English verbs to a NAO robot. The children had to guide the robot's arm to act out the target verbs, e.g. brushing teeth. In a comprehension post-test the children answered correctly more often with words they had taught the robot than those learnt during a regular verb-learning game. While the robot only learned from 'Direct' teaching, where the child was guiding the motion of the robot, there was a high frequency of verbal teaching using English.

We can see that there are many instances where RALL is able to assist in teaching an L2 to students. Many of these show a reduction in FLA and increase in confidence and willingness to learn in the students. In all these cases, however, they use the robot to teach, whether directly in the role of teacher or acting as a care receiver or assistant. Robots were not used in assessment, and in most cases the tests performed were aimed at measuring the comprehension of the L2 words that were being taught. We want to explore the possibility of using a robot to assess the L2 production of children. Due to the reported reductions in anxiety and increase in confidence when using a robot, we may see an increase in the amount of production.

3 STUDY DESIGN

This study was conducted at a local school with English-speaking 5- to 6-year-old children. We decided to teach spatial language, more specifically spatial prepositions, because while those concepts are more abstract than physical objects, we can still represent them using images. Spatial language itself is also particularly challenging to L2 learners as the meaning can often differ depending on context and the referent. Every morning, five children were randomly selected to participate in the study for that day and assigned a condition, balanced across gender. These five children were first given a French lesson before playing our production quiz game on an interactive table [3] individually throughout the rest of the day (Figure 1). An agent (robot or experimenter depending on our condition) is placed opposite to the child and gives instructions and encouragement to the children. The interactive table displays an image of a teddy bear and a chair. The child would have to use one of the French words taught to describe the position of the bear relative to the chair.

As well as the teacher three experimenters were involved in the study:

(1) *Lead Experimenter* - The lead experimenter acted as the interaction point for the children outside of the one to one sessions. Either the lead experimenter or the wizard was required to be in the presence of the child while outside their classroom. The lead experimenter was certified in the children's health and well being, and was there to ensure the health and safety of the children as required by the school.
(2) *Wizard Experimenter* - The wizard experimenter controlled the robot remotely via a laptop interface. The wizard experimenter was also certified in the children's health and well being, but had minimal interaction with the children so as to minimise interference during the study.
(3) *Blind Experimenter* - The blind experimenter facilitated the interactions before the main study began, provided the comprehension test and acted as the agent in the child-human condition. The blind experimenter was unaware of the purpose of the study to reduce influencing the outcome.

3.1 Hypothesis

With our study we wanted to test the following hypothesis:

H The presence of a robot will allow children to produce more spatial words verbally in an L2 than when working with a human experimenter.

3.2 Teaching

The children were taught five French words: *Nounours* (Teddy Bear), *chaise* (chair), *devant* (in front of), *sur* (on), *sous* (under). Of these, the first two were supporting words and the last three were the target words for the study. The content of the lesson was created and taught by a professional French teacher, with a goal of enabling the children to produce these words after one lesson. We decided to use a professional teacher as we did not want a robot teacher that would also influence our results. It has also been shown that human teachers can still outperform a robot teacher [7]. The lead experimenter acted as a teacher's assistant. The children were taught in groups of five. The lesson was designed to last 30 minutes.

The teacher started the lesson by introducing the children to the support words. At all stages the children were encouraged to repeat any French words they heard. The children were taught a song that used the three target words and hand gestures to go along with them. After singing, the children would position themselves relative to the chair based on the words announced by the teacher. The children were then each given a teddy bear and repeated the process with the bear. The children then played a game of 'Telephone'. In this game one child was first given one of the target words, and each child would whisper the word to the next child down the line until the last child. The last child would announce to the rest of the group the word they heard. The game was repeated several times with the children re-organised into a different order so that the announcing child changed each time. This was followed by a game of 'Corners'. In each corner of the lesson area, a teddy was placed in a position relative to a chair that referred to one of the target words. The children were then encouraged to sing and move around until the teacher would stop them, and say one of the target words. The children then had to move to the relevant corner and say the word three times. Variants of this game were then played in teams with the chairs lined up, and then individually. Finally each child was told to say one of the target words and then go stand by the correct chair. The lesson wrapped up with one more repetition of the song they had been taught near the beginning.

During the interaction we also established any prior knowledge in the target language. They were split into the following categories:

(1) *No Exposure* - The children have not been exposed to any French, other than potentially those used in popular culture e.g. *C'est la vie.*
(2) *Beginner* - The child has potentially received some lessons in French and knows simple phrases that do not include our target words e.g. *Je m'appelle John.*
(3) *Intermediate* - The child has knowledge of French, including our target words.
(4) *Advanced* - The child has an intricate knowledge of French, and is able to produce words with a high capability or are fluent.

Children of intermediate or advanced knowledge were excluded from the data analysis. 25 children took part in our study of which three were excluded from the analysis of results, leaving 22 children.

3.3 Individual Interactions

Upon completing another familiarity task and a 10 minute activity with the robot–that required the child to describe the position of objects to the robot in English–a comprehension test was administered by a blind experimenter who was unaware of the purpose of the study (Figure 2). This served as a small refresher of what the children had learned earlier in the day, as well as allows us to establish a baseline for the efficacy of the lesson. For the comprehension test there were 6 sheets with 3 images each (representing the 3 target words), placed on the left, in the centre or on the right. Together, the 6 sheets covered all possible permutations of the 3 target words (*devant*, *sur*, *sous*) with each of the 3 positions. The images were similar but not the same as the ones used for the production quiz questions. For each sheet the experimenter asked the child to point at the picture that matches the statement (see below). If the

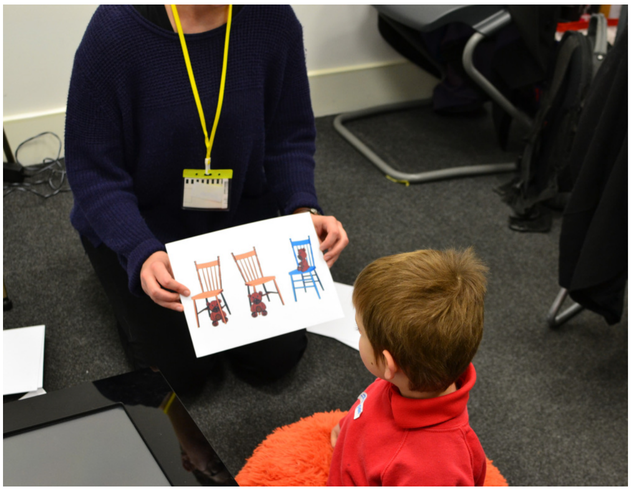


Figure 2: A child being administered the comprehension test before moving onto the main production quiz.

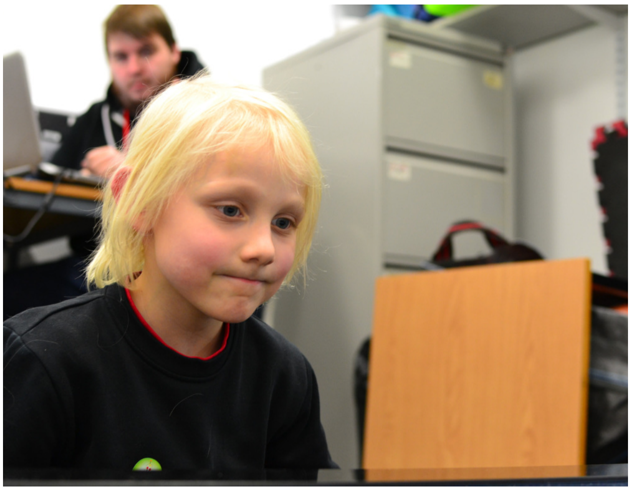


Figure 3: The 'wizard' experimenter was positioned behind the child to minimise interaction between them.

child pointed to the wrong picture they were allowed to try again until they pointed to the correct image. We repeated each target word twice to account for guessing and to ensure they weren't just picking based on location on the question sheet. The statements and their order were the same for every child:

(1) Le nounours est sous la chaise.
(2) Le nounours est devant la chaise.
(3) Le nounours est sur la chaise.
(4) Le nounours est devant la chaise.
(5) Le nounours est sur la chaise.
(6) Le nounours est sous la chaise.

The child then played the production quiz with either the robot or the blind experimenter based on the group they were in (child-robot or child-human). In both conditions, the production quiz was displayed on the sandtray. The robot was controlled through a Wizard-of-Oz interface, with the 'wizard' sat behind the child, out of sight, so as to minimise effects on the child (Figure 3). The rules of the game were explained by the agent (blind experimenter or robot). The child was sat in front of the sandtray upon which the production quiz game was displayed. The agent sat opposite the child. The sandtray displayed an image of the teddy bear in a position relative to the chair, and the agent or child must answer "Où est le nounours?" (Where is the teddy bear?). The agent was to give the answer in the form "sur/sous/devant la chaise", but any answer given by the child that included one of the target words 'sur', 'sous' or 'devant' was accepted. Each correct answer scored a point. If either the question was answered correctly or both the child and the agent answered incorrectly then the production quiz moved onto the next question. If the child did not answer after a short period then the agent would give encouragement in proceeding levels:

(1) Encourage the child to guess e.g. "Just have a guess".
(2) Targeted encouragement, such as asking them to remember the lesson from the morning.
(3) The agent will attempt the question.
    - If the child was ahead on points then the agent (adult/robot) would answer correctly so as to keep up an appearance of a challenging opponent in the game.
    - If the child was level or behind the agent (adult/robot) then the agent would answer incorrectly to demonstrate a willingness to answer even if wrong.

If the child still did not have a guess after all stages then the game proceeded as if they had answered incorrectly. The agent began the production quiz after explaining how to play by answering the first question correctly. There were nine subsequent questions which we expected the child to answer, three for each target word.

4 RESULTS

4.1 Participants

25 children took part in our study of which three were excluded from our analysis of results leaving us with 22 children. 11 Children were in the Human Condition (4 Female) and 11 in the Robot Condition (6 Female). There were 11 5 year olds (6 Female) and 11 6 year olds (4 Female). Of these children two had an L1 other than English (1 Female), but their English level was high enough to still participate.

4.2 Comprehension

We scored the comprehension test by taking the maximum attempts per question (3) and subtracting the number of attempts they took to get the correct answer. This meant each question was scored between 0 and 2, giving a maximum possible score of 12 on the comprehension test. The mean score for the comprehension test was 8.5 (SD=1.92). In the Human condition the children averaged 8.27 (SD=2.20) at the comprehension test while in the Robot condition the children averaged 8.72 (SD=1.68). Using a Welch Two Sample t-test, no significant difference between the two conditions was found (t= 0.55, df =18.72 p=0.59). This shows that the groups between our two conditions were roughly equal in ability before beginning the

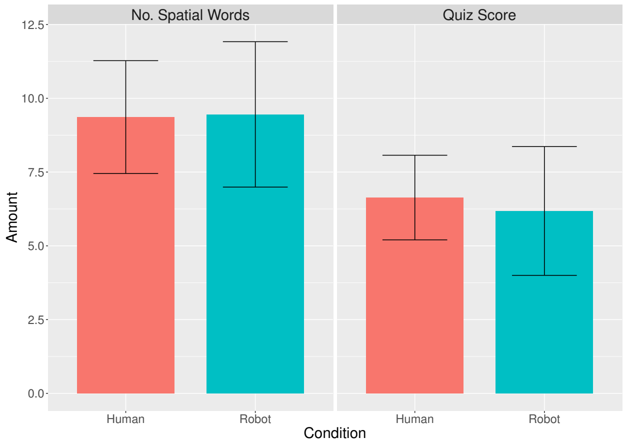


Figure 4: Analysis of L2 spatial words used during the production quiz. Left: spatial words used without additional prompting to attempt the question; right: number of correct words said by the children during the production quiz. In both cases no significant difference was found between the robot and adult conditions. Error bars are showing the standard deviation.

production quiz. The scores remained consistent throughout the test, with no learning effect seen when the first half and the second half of the comprehension test were compared (first half: mean=4.5, SD=1.26; second half: mean=4 SD=0.93; t=1.50, df = 38.51, p=0.14).

4.3 Production

Children in the child-human condition scored M=6.64 (SD=1.43) out of 9 on the production quiz and M=6.18 (SD=2.18) in the child-robot condition. Using a Welch Two Sample t-test no significant difference between the two conditions was found (t=-0.58, df =17.27, p=0.57).

We also analysed the total number of spatial vocabulary used in L2 (Figure 4). Due to a break in protocol, children were sometimes prompted to attempt a question again instead of moving on in the production quiz. As such our analysis is on words used without being prompted for an additional attempt. In the Robot condition, the children averaged M=9.45 (SD=2.46) spatial words, compared to M=9.36 (SD=1.91) in the Human condition. Using a Welch Two Sample t-test no significant difference was found (t=0.10, df=18.4, p=0.92).

Finally we analysed the amount and level of encouragement given (see levels in Section 3.3). While encoding encouragement given to the children we added a fourth level for analysis of the results:

(4) Encouragement is given that changes or disrupts the task, e.g. telling the child that the current question is the same as a previous one.

The mean amount of encouragement given was M=12.36 (SD=7.46) in the Human condition and M=13.09 (SD=7.78) in the Robot condition. No significant difference was found between the conditions (p=0.83). However we see a significant difference in the average maximum level of encouragement per question across the two conditions (Robot: M=1.12, SD=0.57. Adult: M=2.09, SD=1.09, p=0.02). This is strongly influenced by the amount of level 4 encouragement given by the adult, of which we see 33 instances across 10 children. We see a significant difference between the average amount of level 4 encouragement given per child between the amount given in the first half of the study compared to the second showing an increase in deviation from the protocol over time (First Half: M=1.25, SD=.0.88. Second Half: M=4.25, SD=2.64, p=0.04).

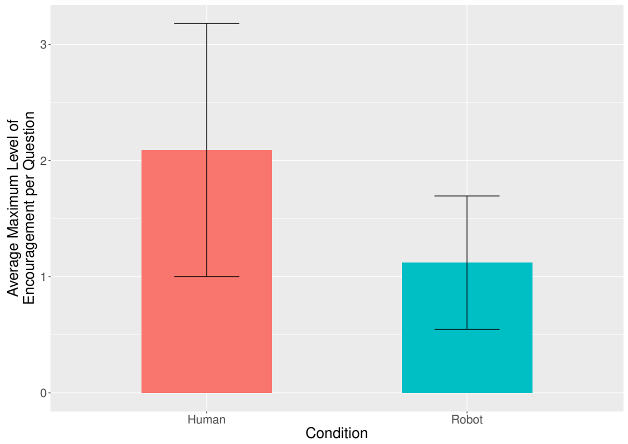


Figure 5: Analysis between participants of the average maximum level of encouragement reached across conditions. A significant difference is seen between the two conditions, Human and Robot. Error bars are showing the standard deviation.

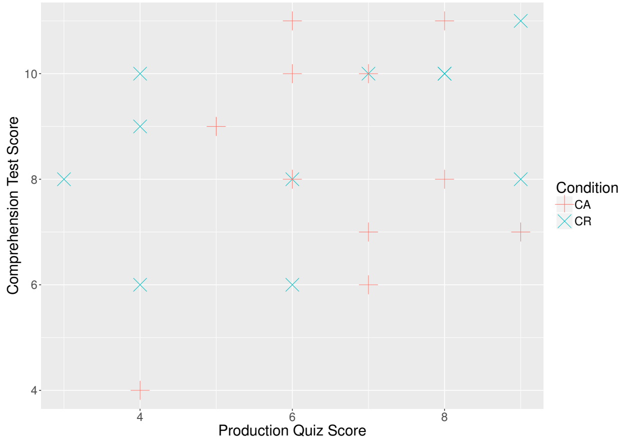


Figure 6: A comparison between the score in the production quiz and the score on the comprehension test. No significant correlation was found.

4.4 Comprehension and Production

The data we collected also provided us with an opportunity to test the predictions of Laufer et al. [9], a key foundation for our research.

By looking at the children's scores on comprehension (passive recognition) and production (active recall) we should be able to see evidence of a hierarchy, where comprehension is required for production.

Across both conditions the children had an average score on the production quiz of 6.41 (SD=1.82) out of 9 and is significantly above chance (p=0.03). A positive but non-significant correlation was found between the comprehension test score and their production quiz score (Pearson's r=0.29, p=0.19). The lack of a significant correlation suggests that abilities in comprehension and production are not directly related.

We marked a child as having achieved comprehension on a particular word if they required less than four attempts across the two relevant questions in the comprehension test. For example if we were looking at whether a child could comprehend the word 'sur' we would look at the number of attempts they took for questions three and five. If a child takes two attempts on question three and one attempt on question five their total number of attempts for 'sur' would be three. We would mark this child as being able to comprehend 'sur'. We marked a child as being able to produce a word if they scored at least two points in the production quiz on the three relevant questions. Using Guttman's Coefficient of Reproducibility (reported in Table 1), we were unable to find a hierarchy. A hierarchy would show that comprehension is needed for production. Guttman's Coefficient measures whether such a hierarchy exists based on the number of deviations from that hierarchy. A coefficient of over 0.9 is expected to display such a hierarchy.

	Sur	Sous	Devant
No. Deviations	5	3	4
Guttman's Coefficient λ_4	0.11	0.57	0.56

Table 1: Table detailing the number of deviations from the expected hierarchy and the Guttman's Coefficient of reproducibility. In the case of all three words, we fail to meet the reliability expectation of 0.9

5 DISCUSSION

5.1 Effectiveness of the robot to support L2 production

While this study does not show statistical improvement to a child's ability to produce by using a robot over a person, it does show a similar performance in this task, with no significant difference between the two conditions being found. It may still be desirable to use a robot to allow standardization and automation of assessment. With a minimal amount of support being provided by an agent, only a narrow set of phrases can be given – otherwise the nature of the task could be changed from production. This can make interactions very repetitive for the assessor. Though the scores were higher than expected it still proved to be a challenging task for the children. With the minimal amount of support available to an experimenter it could be emotionally stressful to be unable to intervene when a child is finding the task difficult.

The scores from the production quiz are higher than we expected. From the literature we expected L2 production to be difficult for the children, and our expert tutor believed that it would take two to three sessions for most children to produce at all. The observed prowess of the children may be partially explained by the design of the lessons, directly aimed at encouraging the children to produce the target words for this study. It should be noted that most productions were only single words. Only two children produced any of the support words (*nounours* – teddy bear, and *chaise* – chair).

Several factors may contribute to the high performance of the experimenter. Even within the context of a limited set of responses a person is able to provide much better cues and encouragement based on reading the child. These kind of social skills are still a gold standard to which robotics researchers strive. Though this experiment was conducted using a 'wizard', their position and the time delay in actions for the robot prevented this fine grained social interaction. Some of the cues provided by the experimenter were not programmed into the robot but should be added into its repertoire

1. *Direct phonetic cues* - Giving part of the word e.g. the starting s.
2. *Indirect phonetic cues* - Giving clues to the word about how it sounds e.g. "It's the one with a strange sound in it"
3. *Rhythmic cues* - Giving the syllables of the word e.g. "Duh-dum". This may work well for the small target vocabulary, like ours, where this could refer to a single word, but may be less effective in larger vocabularies.
4. *Gestural cues* - Movements with the hands that mimic gestures used by the teacher in the lesson.

Despite the more limited social skills of this implementation of the robot, it still achieved a similar performance level to a person. This may be the expected reduction of anxiety, that previous research has shown, balancing the limited social behaviours.

However we also saw a large amount of encouragement given to the children by the blind experimenter that was outside of the original protocol, that could be deemed to have affected the scores of the children in an undesirable way. While in the first half the amount of these encouragements by the experimenter remained low, there was a sharp increase in the latter half. This could be caused by forgetting the protocol over the days of the study or just growing more lax in its use, or even the emotional stress that is put on a person by the children's difficulties.

The presence of a wizard in the room may also have been a contributing factor. The presence of a person, even when not in view, may have prevented the robot from reducing anxiety as much as it could have done, as the child might be aware someone else is listening in. We minimized the affect of the wizard by ensuring there was no reason for them to interact with the children either before the study. Analysis of the videos showed that the majority of children never turned towards the wizard at any point during the study, and focused on the robot. So we believe the impact of the wizard's presence was minimal.

Finally, it must be noted that the school where we performed the study cultivated a much friendlier relationship between adults in the school and the students than is typically seen. This may have made the children feel more comfortable and confident in the presence of our experimenter, reducing anxiety. Future work will focus on broadening this study to multiple schools to see whether our results can be replicated in different settings.

5.2 Relative difficulty of comprehension versus production

The lack of correlation shown between the production quiz score and the number of attempts on the comprehension test (Figure 6) shows that there was no direct relation between comprehension and production vocabularies. However when we look at the possibility of a hierarchy from comprehension to production we do not find evidence to support a hierarchy. This could have had several causes. While we were hoping to find support within our data, we were not directly testing for this hierarchy. Laufer et al. [9] looked at students 16 years and older at high school and university who had been studying their L2 as part of a national curriculum for between 6 to 9 years. Ours is based on a single lesson focused entirely on being able to say the target words. The younger children in our study may also have been more receptive to learning words productively, as they are still increasing their phonological vocabulary. These skills have been shown to have a correlation with word vocabulary [6]. These factors could account for an increase in deviations from the previously established hierarchy.

6 CONCLUSION

We hypothesized that a robot could surpass human performance in encouraging the production of spatial language: this hypothesis is not supported by our study; however, the robot and the facilitator's performance were very similar, with no significant difference between the two conditions being found. This was despite the greater social ability of the human experimenter. This may be explained by the previous research that shows that robots can make people less anxious in foreign language learning scenarios. Future work expanding the robot's social ability may improve the robot's ability to assess and support a student's learning.

Measuring the production skills of a child at this level is a repetitive and lengthy task. An autonomous robot that is able to measure the production level of a child could be used as a tool to alleviate these factors, enabling more accurate data collection for both research and assessment purposes. Currently we are planning on expanding this work to more schools while increasing the social skills of the robot.

7 ACKNOWLEDGEMENTS

This work was supported by the EU H2020 L2TOR project (grant 688014). The authors would also like to thank the teacher, who wished to remain anonymous, who provided the French lessons for the children. All statistics and graphs were obtained using R [15].

REFERENCES

[1] Minoo Alemi. 2016. General Impacts of Integrating Advanced and Modern Technologies on Teaching English as a Foreign Language. *International Journal on Integrating Technology in Education* 5, 1 (2016), 13–26.

[2] M Alemi, A Meghdari, and M Ghazisaedy. 2015. The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition. *International Journal of Social Robotics* (2015), 1–13.

[3] Paul Baxter, Rachel Wood, and Tony Belpaeme. 2012. A touchscreen-based'sandtray'to facilitate, mediate and contextualise human-robot social interaction. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 105–106.

[4] Larry Fenson, Philip S. Dale, Steven Reznick, Donna J. Thal, Elizabeth Bates, Jeff Hartung, Stephen J. Pethick, and Judy Reilly. 1993. *The MacArthur Communicative Development Inventories: UserâĂŹs guide and technical manual.* San Diego, CA: Singular Publishing.

[5] Morrison F. Gardner. 1990. *Expressive One-Word Picture Vocabulary Test - Revised.* Novato, CA: Academic Therapy.

[6] Susan E Gathercole and Alan D Baddeley. 1989. Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of memory and language* 28, 2 (1989), 200–213.

[7] James Kennedy, Paul Baxter, Emmanuel Senft, and Tony Belpaeme. 2016. Heart vs hard drive: children learn more from a human tutor than a social robot. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 451–452.

[8] Batia Laufer. 1998. The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics* 19, 2 (1998), 255–271.

[9] Batia Laufer and Zahava Goldstein. 2004. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning* 54, 3 (2004), 399–436.

[10] Batia Laufer and Paul Nation. 1999. A vocabulary-size test of controlled productive ability. *Language Testing* 16, 1 (1999), 33–51.

[11] Batia Laufer and T. Sima Paribakht. 1998. The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning* 48, 3 (1998), 365–391.

[12] Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, Gary Geunbae Lee, Seongdae Sagong, and Munsang Kim. 2011. On the effectiveness of robot-assisted language learning. *ReCALL* 23, 01 (2011), 25–58.

[13] Didier Maillat. 2010. The pragmatics of L2 in CLIL. Language use and language learning in CLIL classrooms. *Language Use and Language Learning in CLIL Classrooms* (2010), 39–58.

[14] Jan-Arjen Mondria and Boukje Wiersma. 2004. Receptive, productive, and receptive + productive L2 vocabulary learning: What difference does it make? In *Vocabulary in a Second Language: Selection, Acquisition and Testing*, Paul Bogaards and Batia Laufer (Eds.). John Benjamins Publishers, 79–100.

[15] R Core Team. 2017. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[16] Fumihide Tanaka and Shizuko Matsuzoe. 2012. Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction* 1, 1 (2012).

[17] E. H. Wiig, W. Secord, and E. Semel. 1992. *CELF-Preschool: Clinical Evaluation of Language Fundamentals - Preschool.* New York: Psychological Corp.

[18] Kathleen Williams. 1997. *Expressive Vocabulary Test.* Minnesota: American Guidance Service.

Playing Charades with a Robot: Collecting a Large Dataset of Human Gestures Through HRI

Jan de Wit*, Bram Willemsen, Mirjam de Haas, Emiel Krahmer, Paul Vogt,
Marije Merckens, Reinjet Oostdijk, Chani Savelberg, Sabine Verdult and Pieter Wolfert
Tilburg center for Cognition and Communication, Tilburg University
Tilburg, the Netherlands
Email: *j.m.s.dewit@uvt.nl

***Abstract*—This work documents a playful human-robot interaction, in the form of a game of charades, through which a humanoid robot is able to learn how to produce and recognize gestures by interacting with human participants. We describe an extensive dataset of gesture recordings, which can be used for future research into gestures, specifically for human-robot interaction applications.**

***Index Terms*—Robot learning, Human-robot interaction, Gesture recognition, Robot motion**

I. Introduction

The ability to produce and recognize non-verbal communication, such as gestures, facilitates understanding between humans and robots, and results in more engaging interactions [1]. Previous work [2] has also shown that a robot's use of iconic gestures [3] is beneficial to second language learning. By enabling the robot to learn these gestures from demonstration [4], we avoid the need to manually design and program them, thereby removing the influence of the designer's frame of reference. The resulting motions could potentially be perceived as more human-like, because they are based on recordings of human gestures that are automatically mapped onto the robot. In this work, we present a dataset of recorded gestures for 35 different objects, which was gathered through a game of charades with a robot.

II. Approach

A. Procedure

After completing a practice round, the robot started the game by performing a gesture from its set of examples, previously recorded from other participants. The participant was then shown a picture of the item that the robot tried to enact, along with three incorrect answers, on the tablet (see Figure 1, left). If the participant guessed incorrectly, the robot performed a gesture for the same object once more for another guess. Then, the roles were reversed and the participant was shown an object on the screen, which they then described using an upper-body gesture (Figure 1, right). The robot tried to recognize the object that was portrayed, and if guessed incorrectly the participant was asked to perform a gesture for the object again for a second attempt. To provide additional insight into the robot's confidence when guessing, the participant was shown a top five of answer candidates, from which the robot picked the top one for its guess. Each game session lasted five rounds of the robot and participant taking turns guessing, covering ten objects — five performed by the robot, five by the participant — out of a total set of 35, which included animals, static objects (furniture, buildings), tools (e.g., cup, book, toothbrush), musical instruments, and vehicles.

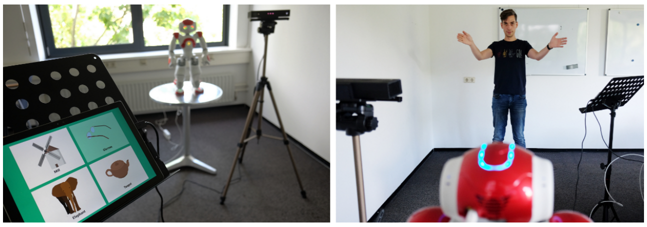


Fig. 1. Left: The participant correctly guesses a gesture performed by the robot. Right: The participant is performing a gesture for the robot to guess (icon designed by Freepik, temporarily added to preserve author anonymity).

B. Implementation

Gesture recognition was implemented by extracting the *gist* of the gesture, inspired by the work of Cabrera and Wachs [5]. This gist was then compared to the complete set of previously recorded gestures using a k-nearest neighbors approach to find the object that was most likely depicted by the current gesture. Hierarchical clustering was used to group similar gestures for each object, and after the participant guessed an object, the weights of these clusters and individual gestures within clusters were increased or lowered based on whether the answer was correct or not. When choosing a gesture to perform, the robot would either explore a new sample (40%), or exploit the cluster and sample with the highest weight (60%). Previously recorded gestures were mapped to the robot's accepted input format for performing motions by calculating the various joint angles that the NAO robot accepts from the joint positions of the participant that were recorded by the Kinect camera. Three gestures for each of the 35 objects were performed by the researcher and added to the system as an initial set for recognition and production. The system was deployed, with an identical setup, at two locations: a science museum that is mostly visited by children and teenagers, with

This research was funded by the EU H2020 L2TOR project (grant 688014).

their parents, and a music festival where most visitors were adults. All recorded data were cleared between the two events, so that the robot would have to start learning from scratch again.

III. Descriptives

The system ran for fourteen days at the science museum, and for three days at the music festival. Table I shows the demographics and number of gestures gathered from each location.

TABLE I
Description of Datasets

	Science museum	**Music festival**
Participants	294	116
Gender	147 Male 141 Female 6 Unknown	49 Male 67 Female
Average age (years)	12.8 (*SD* = 10.7) 10 unknown	28.3 (*SD* = 8.7) 2 unknown
Countries	26	4 (1 unknown)
Number of gestures	2,524	1,000

The recorded gestures were stored in the form of a CSV file containing the 3D coordinates of the participant's tracked joint positions, sampled at approximately 30 frames per second from the Kinect camera, as well as a movie file containing a 2D render of the gesture (Figure 2). Furthermore, gestures can be linked to participants and their demographic information by a unique identifier.

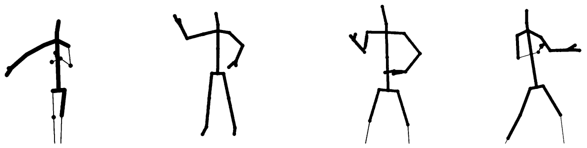


Fig. 2. Four examples of recorded gestures for 'guitar' — first and second are by children, third and fourth by adults.

A. Recognition and Production Performance

After both experiments had finished, we analyzed how the robot's gesture recognition rate developed through time. The results from the science museum are shown in Figure 3. At the music festival, the recognition rate started at 16.37% on the first day, followed by 23.36% and 23.24% on the second and third days. In all cases, the robot performed well above chance (which was approximately 3%). The comprehensibility of the robot's gestures was measured by looking at the number of times participants managed to guess correctly. Figure 3 presents the results from the science museum. During the first day of the music festival, participants managed to guess correctly 50.31% of the time, followed by 51.65% on day two and 50.65% on the last day. This is also above chance (which was 25% for a first attempt, 33% for a second attempt).

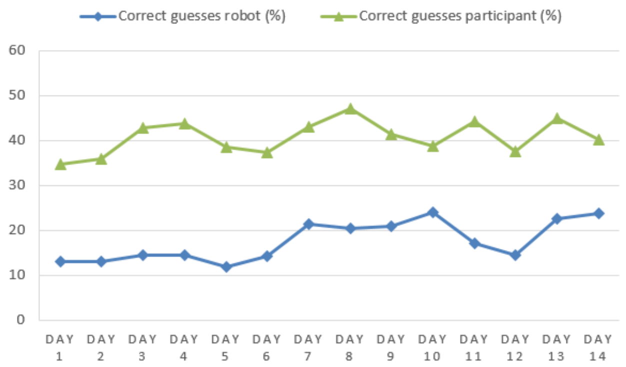


Fig. 3. The robot's and participants' performance (% guessed correctly) during the fourteen days at the science museum.

IV. Conclusion and Discussion

This paper presents an exploratory study where a game of charades was used as a playful method to allow a robot to optimize its own gesture production and recognition abilities. At the same time, an extensive and varied dataset was recorded, to allow future research into gestures, with applications in the field of HRI. We intend to conduct further analyses on the recorded gestures (e.g., which strategies were used, whether these changed between first and second attempts, differences between participant groups), and aim to further improve the robot's ability to recognize and produce gestures. It is difficult to interpret the gesture recognition performance of the system, because existing research tends to work with a smaller set of concepts, and often focuses on detecting a certain predefined gesture, rather than allowing the person performing it to decide on a strategy themselves. However, it does appear that the performance flattens out with a relatively sparse set of data, which can be seen as an indication that we have not reached the maximum potential yet. We would be interested in measuring human performance on recognizing the gestures, to get an idea of the gold standard. The dataset of gesture recordings, as well as the source code of the system will be made publicly available after our further analyses are complete.

References

[1] P. Bremner, A. G. Pipe, C. Melhuish, M. Fraser, and S. Subramanian, "The effects of robot-performed co-verbal gesture on listener behaviour," in *Proceedings of the 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*. IEEE, 2011, pp. 458–465.

[2] J. de Wit, T. Schodde, B. Willemsen, K. Bergmann, M. de Haas, S. Kopp, E. Krahmer, and P. Vogt, "The effect of a robot's gestures and adaptive tutoring on children's acquisition of second language vocabularies," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2018, pp. 50–58.

[3] D. McNeill, "So you think gestures are nonverbal?" *Psychological Review*, vol. 92, no. 3, pp. 350–371, 1985.

[4] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[5] M. E. Cabrera and J. P. Wachs, "A human-centered approach to one-shot gesture learning," *Frontiers in Robotics and AI*, vol. 4, p. 8, 2017.