Second Language Tutoring using Social Robots: A Large-Scale Study

Anonymous Placeholder
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

***Abstract*—We present a large-scale study of a series of seven lessons designed to help young children learn English vocabulary as a foreign language using a social robot. The experiment was designed to investigate 1) the effectiveness of a social robot teaching children new words over the course of multiple interactions, 2) the added benefit of a robot's iconic gestures on word learning and retention, and 3) the effect of learning from a robot tutor versus learning from a tablet application. For reasons of transparency, the study's research questions, hypotheses and methods were preregistered. With a sample size of 192 children, our study was statistically well-powered. Our findings demonstrate that children are able to acquire and retain English vocabulary words taught by a robot tutor to a similar extent as when they are taught by a tablet application. In addition, we found no direct benefit of a robot's iconic gestures.**

***Index Terms*—Robots for learning; Second language tutoring; Child-Robot Interaction; Long-term interaction; Gesture**

I. Introduction

Social robots have shown considerable promise as teaching-aids in education, where they can be deployed to support learning of constrained topics [1]–[3]. Next to STEM topics, (second) language tutoring is seen as an area for which robots can offer effective educational support [4]–[7]. Robots not only hold the promise of a more effective one-to-one delivery of tutoring, for which there is little time in current educational practice, they also promote social behaviours which are conducive to learning, such as sustained attention and compliance. One assumption for why social robots can be good language tutors, especially for younger children, is that robots have the ability to physically interact with children in the real world in a semi-naturalistic manner, both verbally and non-verbally. However, it is still unclear to what extent robots can be effective tutors of a second language (L2), and how to best design effective robot language tutors. We believe that one reason for this is that current studies are statistically underpowered and often glean results from only a single interaction session. In this study, we address these issues in a large-scale study in which preschool children learn words in an L2 over multiple one-on-one tutoring sessions.

Many studies are often small-scale and short-term, involving typically one interaction session with a relatively small sample size [8], [9]. The reason for this being that developing and carrying out human-robot interaction (HRI) experiments is time-consuming and costly, especially for long-term interaction studies [10]. Results from short-term studies may be severely biased, as learners will not have previously interacted with a robot and the interaction might therefore be influenced by the "novelty effect". Learners' attention might be affected; instead of attending to the task at hand, learners may focus predominantly on the robot and its behaviour instead. First interactions also involve some anxiety or excitement about the encounter, which can reasonably be expected to influence learning outcomes. As such, long-term studies are essential to investigate the effect of interacting with a robot on multiple occasions, especially since many studies have shown that the novelty effect rapidly wears off (see [11] for an overview). Long-term studies are particularly critical in educational robots, because learning a particular skill, such as speaking and understanding an L2, requires repetition and time [12].

Few studies have investigated the effect of robots in multiple lessons on language learning [5], [7], [13], [14], with mixed results. For instance, Kanda and colleagues [5] did not observe a clear learning effect in their two week field trial, except that children who interacted longer with the robot during the second week scored higher on the English post-test. However, it could be that these children interacted more often with the robot, because they were more proficient in English. Kanda et al.'s study revealed that most children lost interest in the robot, possibly because they had difficulties understanding the robot, but also because the novelty effect may have worn off [5]. On the other hand, studies by Lee and colleagues [13] and Tanaka

This project has received funding – details to be included in the final version

and Matsuzoe [14] have demonstrated that children can learn a limited L2 vocabulary from a robot over the course of multiple interactions.

These long-term studies were, however, very exploratory in nature due to the small sample sizes (18-21 students) and only one experimental condition, as a result of which they can only offer a "proof of concept". To investigate, for instance, the added value of using a robot or a particular interaction strategy, multiple conditions need to be investigated using a statistically well-powered sample size. Those studies that increase the sample size, tend to either have only a single session [5] or have only one condition [15].

So, to what extent are robots effective L2 tutors? And if they are, are they more effective than other digital (screen-based) tutors, and why? A good argument for why robots could be effective tutors comes from the notion of embodied cognition. Human language use is grounded in our interactions with other language users and our interactions with the physical world [16]. Compared to other screen-based technologies, the interactions with a physical robot provide such grounding and are situated in a three-dimensional, tangible world [17]. The physicality of the interaction allows for a true implementation of the embodied cognition paradigm [18], which holds that our cognition is anchored to our bodily experiences with the real world.

One of the features in which the physicality of the interaction can manifest itself is by having robots interact multimodally. In particular, it has been suggested that robots' ability to produce gestures can have an added value for L2 learning. In gesture research, one often distinguishes deictic gestures (such as pointing or showing) from iconic gestures (where the shape of the gesture has some physical similarity to its referent) [19]. Both forms of gestures can have a positive effect on L2 learning. Deictic gestures help to establish joint attention, which in turn benefits the learning of word-meaning mappings [20]. Iconic gestures produced by tutors can also have a positive effect on vocabulary learning in children [21] and in adults [22], [23], and even when the gestures are produced by robots [15]. The exact reason why gestures can be beneficial is not entirely clear, but it may be that they can help identify the meaning of words [24] or perhaps indirectly activate associations in the motor cortex that simulate (or even activate) the production of gestures by the learner, which can help to strengthen the association between word and meaning [18].

In the current study, we investigate the effect that robots –either using iconic and deictic gestures or only deictic gestures– may have on teaching 5- to 6-years-old children basic vocabulary from a foreign language in a longitudinal study over seven sessions. Moreover, the effect of the robot tutor is compared to a screen-based implementation on a tablet computer. In contrast to many other previous studies, the study is statistically well-powered with a sample size of 192 children. The experiment has four conditions:

1) *Robot with iconic gestures* where the robot supports tutoring using iconic and deictic gestures, and with interactions mediated by a tablet game.
2) *Robot without iconic gestures* where the robot supports tutoring without using iconic gestures, but with deictic gestures, and with interactions mediated by a tablet game.
3) *Tablet-only* without a robot present, but with audio lessons using the robot's voice, and where interactions were mediated by a tablet game.
4) *Control* condition where children danced with the robot but were not exposed to the educational material.

In this paper, we investigate the effect that the different conditions have on learning performance. Based on predictions both from the literature on learning and earlier studies with robot tutors, we formulate 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 control condition.

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

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.

The study's research questions, hypotheses, and methods have been preregistered at AsPredicted.[1] By preregistering all these elements, prior to the data collection, researchers are committed to present their analyses based on what they registered in advance. This ensures transparency and would thus reduce an often used practice of selectively choosing or adapting research questions, hypotheses or methods after the data collection. This does not mean that one cannot explore the data any further, but it urges researchers to at the very least present their study as it was originally designed [25].

In the remainder of this paper, we first outline the lesson plan and the basic interactions we designed between the young learner, robot and tablet. In Section III we will explain our methods. Section IV presents the results, which we discuss in Section V.

II. Lesson series

Lessons were designed to teach English vocabulary to 5- to 6-year-old native Dutch speaking children using the NAO robot as a (nearly) autonomous tutor. All lessons involved one-on-one interactions between robot and child. Since no reliably performing automatic speech recognition for children's speech exists yet [26], the interactions were mediated through a game played on a Microsoft Surface touch-screen tablet computer, which provided visual context. The basic setup used throughout the lessons is shown in Figure 1. In this setup, the child would sit on the floor in front of the tablet (i.e. from the position where the photograph was taken). The NAO robot

[1] See AsPredicted.org –the exact URL with the preregistration has been omitted for anonymity.

was placed in a crouching position in an angle of 90 degrees towards the child, also facing the tablet, which was placed on top of a small box. A video camera placed on a tripod facing the child was used to record the interaction. A second camera was placed from the side to get a more complete overview of the interactions.

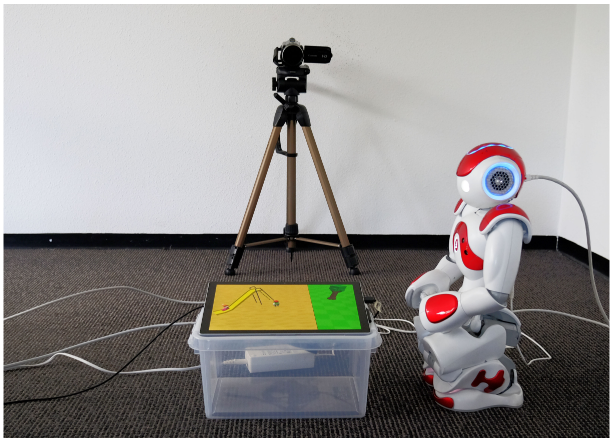


Fig. 1. The basic setup for all lessons.

A. Target words

English target words were selected for two domains in the academic register, which contain words that are typically used at schools. The two domains were mathematics (i.e. words involving numeracy, such as counting words, basic maths and measurement) and space (i.e. words involving spatial components, such as spatial relations, prepositions and action verbs). In addition to the target words, various support words in English, such as animal names (e.g., giraffe, elephant or monkey) or other nouns (e.g., girl, boy, ball), were used to embed the target words in English phrases.

In total 34 words were selected. Selection was based on school curricula, child-language corpora, and age-of-acquisition lists. Target words were selected such that they occurred in school curricula, and that children had already acquired them in their first language. The goal of the intervention was not to teach children new mathematical and spatial concepts, but rather to teach L2 labels for mathematical and spatial concepts that children were already familiar with.

The 34 target words were introduced to the children in 6 lessons each including 5 or 6 words and were recapped in a 7th lesson. Each target word was repeated at least 10 times in the lesson in which it was introduced. In addition, each word was repeated once more in the subsequent lesson, and at least twice in the recap lesson. Words were repeated more often if children required additional feedback. Each lesson was situated in a particular location displayed on the tablet screen, such as a zoo, bakery shop or playground, and focused on teaching target words around a particular theme. Table I shows the settings and target words for the seven lessons.

TABLE I
OVERVIEW OF THE LESSON SERIES.

L	Setting	Target words
1	Zoo	one, two, three, add, more, most
2	Bakery	four, five, take away, fewer, fewest
3	Zoo	big, small, heavy, light, high, low
4	Fruit shop	on, above, below, next to, falling
5	Forest	in front of, behind, walking, running, jumping, flying
6	Playground	left, right, catching, throwing, sliding, climbing
7	Picture book	*all target words*

B. Lesson plan

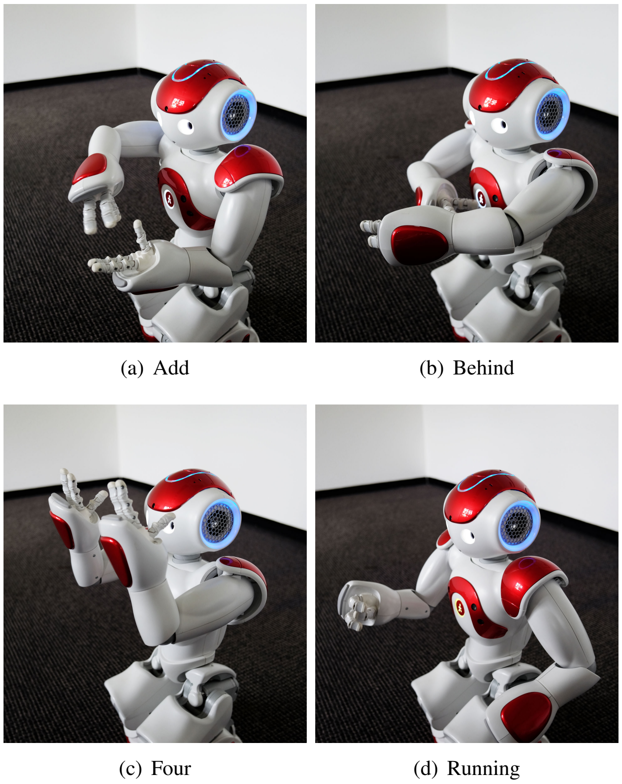
Each of the 6 content lessons consisted of three phases. The first phase was a brief introduction with a personalized greeting, a short reminder of the previous encounter and an introduction of the new location that set the context of the lesson at hand. The second phase was a word modelling phase where the children learned what the target words referred to, while they were named in both Dutch and English together with an example shown on the tablet. Typically, a new target word was introduced in a game-like fashion where the concept appeared on the screen (sometimes in conjunction with one or more support words that were introduced earlier). The robot then provided a comment and the target word in Dutch, and asked the child to touch the target object. The English target word was then first introduced by the tablet through a pre-recorded voice from a native English human female speaker. The robot repeated the word and asked the child to repeat the target word too. Although we aimed for full autonomy, this was the only place where we had to rely on Wizard of Oz (WoZ) to indicate whether the child had said something, because neither automatic speech recognition nor automatic voice activity detection worked sufficiently reliable. Irrespective of what the child had said, if the child had tried to repeat the robot, positive feedback was provided. If the child remained silent, the robot would motivate the child to talk by asking again up to two times. If the child still had not responded verbally, the robot proposed to repeat the word together with the child, and count down from 3 to 1. After that the lesson would indeed proceed irrespective of the child's response.

Let us illustrate the word modelling with an example. In lesson 1 after the support word 'monkey' was introduced, the robot asked the child to put the monkey in its cage (using the tablet). After this was done, the robot continued to say: "In the cage there is now one **monkey**. Let's hear the word for one in English. Touch the monkey in the cage". (Note that in our examples, everything is said in Dutch, except words or phrases written in bold face.) When the child then touched the monkey, a human female voice said "**One monkey**" in native English, after which the robot says: "Ah, one is **one**. Can you say **one**?" And the child was expected to repeat the robot saying 'one'.

After a target word was thus introduced, the robot and child would engage in certain tasks that revolve around the target word. For instance, the child was asked to place 'one', 'two' or 'three' animals in a cage, or 'adding' them. The tablet software monitored whether the child was doing so correctly and the robot provided feedback. The way feedback was provided varied: there were 11 variations of positive feedback phrases, 10 for negative feedback, and 7 for speech-related tasks. Positive feedback was always non-specific (e.g., "Well done!"), but negative feedback incorporated context (e.g., "Nice try, but you need to touch the monkey in the cage. Try again"). All feedback variations were derived from an (unpublished) interview study with student teachers. When children continued to fail a certain task twice in a row, the robot would 'magically' demonstrate how to do this by swiping its arm over the tablet causing the desired action (e.g., placing a monkey in the cage) to occur.

Once all target words were modelled, each lesson would end with a short test in which knowledge of each target word was tested twice in a random order. For each test item, the tablet showed three pictures or animations with familiar objects/actions from that specific lesson, and the child was asked to tap on the relevant picture/animation. During these tests, the robot did not provide any feedback nor gestures to help children. The results of these tests are not analysed within the scope of this paper.

The seventh session was a recap lesson, where children created a picture book. They saw, one by one, the scenes of the six content lessons, and 'stickers' with the objects of these lessons. They placed these 'stickers' on the scenes, while the robot discussed with the children the target words that they were taught during that lesson.


(a) Add

(b) Behind

(c) Four

(d) Running

Fig. 2. Examples of iconic gestures used in this study, photographed from a position where the child would sit. (a) The word "add" is depicted with the right hand as a place holder, and the left hand moving as if it puts something there. (b) The word "behind" is gestured by moving the left hand up and down behind the right hand. (c) The word "four" is depicted by holding both hands up, such that it shows four fingers when viewed from the front. (d) "Running" is gestured by moving both arms back and forth as if the robot is running.

C. *Different conditions*

The content of all seven lessons was exactly the same for all conditions, except the control condition. Differences between the three experimental conditions concerned the modality in which content was presented and the physical presence of the robot.

1) Robot with iconic gestures.: In this condition, the robot would produce an iconic gesture each time it uttered a target word in English. The iconic gestures produced represented the target word in an iconic way. For example, the word "one" was gestured by holding up one hand as a fist; "two" by extending the hand with the back facing the child, so she saw only two fingers; "three" was shown by holding up its hand with the palm facing the child showing all three fingers. "In front of" was shown by moving one hand in front of the other hand; "behind" was gestured by moving one hand behind the other hand. Fig. 2 shows some example gestures. The iconic gestures used in the lessons were designed following an experiment in which several adult participants were asked to depict each target word, and the resulting gestures were tested on clarity using other adults [27].

2) Robot without iconic gestures.: Here, the robot would not produce iconic gestures. However, this does not mean that the robot did not gesture at all in this condition. In both robot conditions, the robot occasionally produces a deictic gesture. Sometimes it would point to the tablet to draw the child's attention to some activity happening there, and sometimes when a child did not respond to an instruction to manipulate something on the tablet, the robot would perform the aforementioned 'magical' demonstration of how to execute the task.

3) Tablet-only.: In this condition, the robot is hidden from the child's view. The robot's voice is directed to come from the tablet's speakers and the information displayed on the tablet is exactly the same as in the two robot conditions. The reason for hiding the robot in a large bag, instead of not using it at all, is that this allowed us to use exactly the same software that runs on the robot. Although some children were disappointed for not interacting with the robot (while their classmates were), none of the children seemed to notice the hidden presence of the robot. To compensate these children, we organised a group session with the robot, similar to the introduction (see next section), after the immediate post-test was administered.

4) Control.: Here, children did not receive a lesson, but instead engaged with the robot in three brief one-on-one sessions. In these sessions, the robot would say something nice and personal in Dutch and then the robot and child would dance a popular Dutch children's song.

III. Methods

A. Participants

A total of 208 children were recruited from 9 different primary schools in the Netherlands. The average age was 5 years and 8 months ($SD = 5$ months) and all children were native speakers of Dutch. To ensure that their prior knowledge of English was not too high, children could only participate if they would not exceed a score of 17 on the English pre-test. Three children were excluded after the pre-test as their score on the English pre-test was higher than 17. The children were pseudo-randomly assigned to one of the four conditions, ensuring an equal gender balance and allowing fewer children in the control condition. During the experiments, 10 children dropped out for various reasons, such as fussing and shyness. Data of additional 3 children was excluded as they missed one lesson ($N = 1$) and/or had received one lesson twice ($N = 2$), due to technical issues. The resulting sample included 192 children. Table II shows how the final set of participants are divided over the four conditions.

Children's legal guardians signed informed consent forms, and the experiment was carried out with approval of our institutional Research Ethics Committees.

B. Materials

1) Pre-tests: Before the tutoring sessions started, we pre-tested the target vocabulary (the 34 English words). In the pre-test, children were presented with each of the English target words, and asked what it means in Dutch (Wat betekent het in het Nederlands?). The test was administered using a laptop computer from which the English words, recorded by a native English female speaker, were presented.

In addition, we tested the following items that are known to influence the children's ability to learn language:

- Dutch vocabulary knowledge (Peabody Picture Vocabulary Test) [28],
- selective attention (visual search task) [29], and
- phonological memory (non-word repetition task) [30].

2) Post-tests: We conducted two post-tests (one immediate post-test, administered maximally 2 days after the final lesson, and one retention test, which took place between 2 and 5 weeks after the 7th lesson). Both post-tests contain three parts:

- translation from English to Dutch,
- translation from Dutch to English, and
- comprehension test of English target words.

TABLE II
OVERVIEW OF THE PARTICIPANTS IN THE EXPERIMENT.

Condition	N	Gender	Avg Age + SD	
		N_b/N_g	(Y;M)	(M)
Iconic gesture	53	30/23	5;8	5
No iconic gesture	54	28/26	5;8	5
Tablet	53	24/29	5;9	5
Control	32	14/18	5;6.8	5

For the two translation tasks all 34 target words were tested using the same procedure as in the pre-test. The comprehension task had the format of a picture selection task in which children were shown three pictures or videos simultaneously and asked to choose the picture or video corresponding to the target word. Target words were thus tested three times, which is a standard way in language learning studies to reduce the bias that may result from guessing. However, since doing this for all 34 target words would take too long, a pseudo-random selection of 18 (53%) of the target words were used, containing all the word categories taught (e.g., counting words, verbs etc.). The total score was the number of trials performed correctly and ranged between zero and 54 (= 18 words x 3 trials per word). If children were to guess the correct answer, they would have a chance of 1/3 to choose the correct answer, so only scores above 18 (=54/3) can be considered as scores above chance level.

During the pre-test and the immediate post-test, additional questions were asked about the children's perception of the robot. The results of these questionnaires are presented in [**anonymous**].

C. Procedure

Approximately one week prior to the first lesson, the children participated in a group session where they were introduced to the robot by one or two experimenters. The robot was introduced as 'Robin the robot' and was framed as a peer who would join the children to learn English. During the introduction, children were given information about the robot to establish common ground and were explained how to interact with the robot. For instance, children were told that Robin the robot has something that looks like a mouth but that does not move when it speaks, and that although the robot has large looking ears, they should speak loud and clearly to its face when addressing the robot. Towards the end of the introduction, the children engaged in a short dance with the robot.

After the introduction session, but prior to the first lesson, a trained researcher administered the pre-tests in a one-on-one session. Children are awarded stars for completing various sections of the test. The pre-test took approximately 40 minutes per child.

For each tutoring session with the robot, children were collected from their classroom and brought to another classroom devoted to the experimental setting. The child was placed in front of the tablet and in a 90 degrees angle with the robot (see Fig. 1) and the researcher would start the lesson. During the first part of the lessons, the researcher would help the child if needed by encouraging her to touch the display or telling her that it is her turn to answer the robot. Otherwise, the researcher would sit somewhere behind the child and operate the wizard to proceed the interaction when the child responded verbally to the robot's request. If the child had to go to the bathroom or if the robot crashed (which happened infrequently), the lesson was paused and would continue after the child or robot was ready again. At the end of each lesson, the child was rewarded a star and brought back to the classroom. The duration the experimental sessions varied per lesson and per condition between 16 and 19 minutes on average; with lesson 7 (the recap lesson) taking longest and lesson 1 being the shortest. Lessons in the iconic gesture condition took the longest, followed by the no iconic gesture condition and the tablet condition. The sessions of the control condition were significantly shorter and only took about 5 minutes per session.

After all 7 lessons were completed, the two post-tests were administered by a trained researcher. As for the pre-tests, the post-tests were administered in one-on-one sessions using paper score sheets. The immediate post-test, which contained some additional materials, took about 40 minutes, while the retention test took 30 minutes.

IV. Results

MANOVA and chi square tests showed that the children in the four conditions did not vary in age, gender, level of Dutch vocabulary, phonological memory, selective attention and level of knowledge of the target words prior to the training. Table III shows the main findings from the different tests. One sample t-tests revealed that children score significantly higher than zero on the pre-test translating English to Dutch ($M = 3.5$ words; $t(191) = 16.25; p < .001$). All other translations tasks from the two post-tests also differ significantly from zero ($ps < .001$). While the scores of the translation tasks increase slightly, these are still much lower than the maximum score that could be achieved (34 words). A series of paired t-tests revealed that the translations from English to Dutch measured in the first post-tests are higher than those measured in the pre-tests for all experimental conditions ($ps < .001$) and for the control condition ($p = .008$). Scores on the comprehension tasks were drastically higher than those of the translation tasks and well above chance (18 words) for all conditions ($ps <$ .001).

TABLE III
The main test results.

Condition / Test	Pre-test	Post-test	Retention
Iconic gesture			
Trans(En-Du)	3.38 (3.07)	7.47 (5.16)	8.15 (5.01)
Trans(Du-En)		6.08 (4.19)	6.57 (4.65)
Comprehension		29.30 (5.80)	30.45 (6.29)
No iconic gesture			
Trans(En-Du)	3.59 (3.14)	7.83 (4.94)	8.02 (4.92)
Trans(Du-En)		6.54 (4.28)	6.44 (4.59)
Comprehension		29.50 (6.13)	30.45 (6.29)
Tablet only			
Trans(En-Du)	3.91 (2.80)	7.70 (4.73)	8.42 (4.75)
Trans(Du-En)		6.49 (4.10)	6.70 (4.29)
Comprehension		29.38 (6.44)	30.17 (6.60)
Control			
Trans(En-Du)	2.81 (2.83)	3.81 (3.21)	4.34 (3.22)
Trans(Du-En)		3.16 (2.27)	3.47 (2.13)
Comprehension		25.03 (6.66)	26 (6.04)

All scores indicate the average number of words correctly translated or comprehended. Minimum scores are 0, maximum scores are 34 for translation and 54 for comprehension. For comprehension, chance level is 18.

To test our hypotheses, we performed a 4 (condition) × 2 (post-tests) MANOVA with the three measures at the two post-tests as dependent variables. The findings showed a main effect of condition ($F(9, 452.8) = 2.16, p = .023, \eta_p^2 = .034$). Post-hoc tests (Bonferroni) showed that children in the experimental conditions scored higher than children in the control condition on all tasks ($ps < .05$), but there were no significant differences between the experimental conditions ($ps > .10$). Also, a main effect of time revealed that scores of the retention test were significantly higher than at the immediate post-test ($F(3, 186) = 5.00, p = .002, \eta_p^2 = .075$), suggesting that newly learned words need time to become consolidated.

Finally, we tested a model where children's level of Dutch receptive vocabulary and phonological memory were entered as control variables. This was done by conducting three multiple regression analyses with the three tasks of the immediate post-test as dependent variables. These analyses revealed, besides the effect of condition already shown in the previous analysis, a main effect of general Dutch receptive vocabulary: children with larger vocabularies learned more English words (βs between .14 and .16, $ps < .05$). Effect sizes are small to medium (R^2 ranges from .09 to .13). No effects of phonological memory and no interaction effects were found. When these analyses were repeated with the tasks of the retention test as dependent variables only a significant main effect of condition was found.

V. Discussion

In this paper, we present a large-scale evaluation study that was conducted in order to investigate to what extent social robots can have an added effect in L2 tutoring for preschool children. We investigated the contribution of the use of iconic gestures in the interaction, we compared two different robot conditions with one in which children received the same input from a tablet computer, and we compared all these conditions to a control group in which children did not receive any language tutoring intervention. This study is unique in many respects: (1) we addressed the need to learn in multiple sessions and at the same time overcome issues concerning the novelty effect by providing Dutch speaking children with 7 lessons in which they were taught a total of 34 English words; (2) this study was statistically well-powered with a total of 192 children participating in one of four conditions; and finally, (3) the experiment's research questions, methods and hypotheses were preregistered to ensure transparency about the way that our study was planned, and the way data were collected and analysed.

To summarise the findings, we find evidence to support hypothesis H1 that children can learn L2 target words from a social robot and that they can remember them better than children who participate in a control condition. This is crucial, as it demonstrates that children can, indeed, effectively learn foreign words from a social robot. We, however, do not find evidence to support hypothesis H2 that children will learn more words and remember them better when learning from a robot than from a tablet only. In fact, the results indicate that children learn equally well from the robot as from the tablet. Consequently, these findings do not demonstrate an added value of using a social robot compared to a tablet computer. Finally, we also do not find evidence to support hypothesis H3 that children will learn more words and remember them better when learning from a robot that produces iconic gestures than from one that does not produce such gestures. Although previous studies on L2 learning have demonstrated a positive effect of iconic gestures on learning L2 words [15], [21], [22], the present study does not confirm this. In the remainder of this section, we will elaborate on these findings.

A. Learning from social robots

While it is within our expectations that children can learn L2 from a social robot over multiple lessons [6], [14], [31], it was crucial that we demonstrated that our implementation was effective at teaching the children new vocabulary. Children in the control condition score higher on the two post-tests than on the pre-test in the English to Dutch translation task, and they also score significantly higher on the retention tests than on the immediate post-tests. This demonstrates that these children, despite not having received any lessons from the robot, learned something. They may have learned from carrying out the tests, but also from talking to the children who did receive one of the experimental conditions, or even from elsewhere (after all, most children also knew some English target words prior to our experiment).

The increase in scores on the English to Dutch translation tasks between the pre-tests and post-tests clearly demonstrate that the children are learning during the lessons. The effects, however, appear relatively small, especially when looking at the scores of the translation tasks, which are around 8 out of 34 in the two post-tests of the experimental conditions. Although this seems low, it is consistent with findings from other studies on second language learning demonstrating low scores on children's production in translation tasks [32]. Translating words from Dutch to English seems even more difficult, yielding scores around 6.5 in all experimental (i.e. non-control) conditions. Comprehension scores are considerably higher, as this task is generally easier. The learner only has to recognize the target word from a small set of pictures or videos, instead of having to retrieve and produce the word without context. Chance selection would yield a score of 18, and in all conditions children perform significantly better than chance, and children in the experimental conditions perform significantly better than in the control condition.

To understand why effects are relatively small, one should first consider what the effect size would have been if the same lessons were delivered by a human tutor. This question is hard to answer as we did not measure this, but it is conceivable that the effect size would have been very similar provided the lessons were exactly the same. In order to develop a systematically controlled experiment, all children received exactly the same lessons, except for the variation between experimental conditions and some individual differences due to the amount of feedback received. So, if a human teacher would stick to the exact script of the lessons, the outcome may have been very similar. However, a skilled human tutor would adapt to the individual needs of each child, and present the materials in different ways, possibly using different strategies, to teach and test the child's vocabulary, and respond appropriately to the child's behaviour. Ideally, a robot tutor can do this too. Technologically it is still quite difficult to achieve personalized adaptation in a autonomous robots, although some studies have demonstrated how a robot could adapt to children's correct and incorrect responses [15], [33]). Question remains, of course, how our findings compare to the effect that can be expected when children learn foreign words from human tutors.

B. Social robots vs touch-screen tablets

For social robots to be accepted as an educational tool in schools, it is necessary to demonstrate that they are –at least– as good as other digital tools, such as touch-screen tablet applications, and preferably better. The results of our experiment demonstrate that children learn more-or-less equally well in the two robot conditions as in the tablet only condition. To appreciate these findings, it is important to understand the similarities between the conditions. All interactions in the two robot conditions are mediated by the tablet, which displays the learning context and records the child's input and responses to the system. So essentially, the children play educational games on the tablet. In the two robot conditions, the robot provides verbal support in the form of instructions, translations, and feedback, as well as non-verbal support in the form of deictic gestures and (in one condition) iconic gestures. In the tablet only condition, the verbal support was exactly the same (the robot's voice was directed through the tablet's speakers), but the non-verbal support was not provided.

Although we believe the non-verbal support could provide essential information that would improve second language learning, the fact that in the tablet condition children could focus their attention solely to the tablet game may have boosted their learning performance. From the experiences of the experimenters, it was obvious that in all conditions the children were primarily engaged with interacting with the tablet as this was where most activity took place. One could argue that in the current set-up, the robot was distracting the children playing their games on the tablet, especially in the non-verbal modality. We are currently analysing children's task engagement and their social engagement with the robot from all videos to investigate how engagement varied over the different conditions. We might find a stronger task engagement in the tablet condition than in the robot conditions, although this need not be true. Having a similar or lower level of task engagement in the tablet condition could also be compensated by the fact that children do not need to shift attention from tablet to robot and back. Duration of the sessions might also have some influence, as children's attention span is limited. However, the average duration of the tablet condition sessions were similar as for the robot without iconic gestures sessions; the duration was considerably shorter compared to the robot with iconic gestures condition.

It is justified to wonder to what extent the tablet is hampering the interaction between child and robot. One could argue that interactions without mediation from the tablet, the robot could be much more effective. We agree with this, and the primary reason for mediating the interactions with the tablet is that we aimed for a fully autonomous system. However, since automatic speech recognition for child speech is notoriously unreliable [26] and automatic object tracking is also very hard to achieve reliably [34], we decided to have the interactions mediated by the tablet. If ASR and object recognition would work flawlessly, different and more natural interactions could have been designed that would have exploited the benefits of the robot's attractiveness and embodiment more strongly than in the current experiment.

Note that although we aimed for full autonomy, we have decided to use a WoZ method to replace automatic voice detection, since a pilot study demonstrated that its poor performance hampered the smoothness of the interactions. The robot would either continue and praise children for having repeated the target word successfully in situations they did not, or the robot would continue to wait for a verbal response whilst the child had already responded (perhaps as a whisper). To keep interactions running sufficiently smooth and allow for children to actually say the words as part of the lesson, we decided to opt for the WoZ, but only for this purpose.

C. Iconic gesturing

Given that research has shown that iconic gestures can help people learn vocabulary in L2 [21], [22], even when supplied by a social robot [15], we expected to see an effect too in this experiment. However, our hypothesis on this issue was not supported. It is unclear why this is the case, but it may be due to the clarity of the gestures. They may not have been clear, despite our best efforts in designing the gestures. We used adults to propose gestures, which were then rated by other adults and children –first as they were produced by adults, second as produced by the robot. The design of the gestures was constrained by the physical limitations of the robot, the sometimes clumsy movement of its limbs and the sometimes ill-chosen viewpoint. For example, while humans tend to count on their fingers one to ten, the NAO robot has only three fingers on each hand, which it can only move simultaneously. The robot can gesture 'two' by by holding out a hand with the back facing the child and the fingers stretched, and 'three' by showing the hand with the palm facing the child. Various combinations of these hand positions allowed us to use iconic gestures for teaching the numbers two to five (see Fig. 2 (c)). However, we did not take into account that the child would see the hands from a 45 degrees angle (Fig. 2), which could have been confusing.

Another reason why iconic gestures may not have yielded the expected effect is that they were shown for all target words each time a word was expressed. This could have been an overkill of gestures that also caused the iconic gesture condition to be substantially slower, and which may have distracted the child too much from the learning task (cf. [35]). It might be more useful to have the robot produce the gesture less frequently and only at functionally more appropriate moments, e.g. only when a word is first introduced and when they need extra feedback.

Finally, it may also be that certain types of iconic gestures work better than others. We are currently analysing the data on an individual word level to see whether certain gestures do have an effect on learning. Moreover, some studies have suggested that the bodily (re-)enactment of gestures (or other activities) can have a positive effect on learning [18]. In our experiment, children were only in later sessions occasionally asked to enact a certain concept (e.g., running). We are also currently analysing to what extent children re-enact the gestures and whether this has a positive effect on their learning outcomes. If that is the case, it might be more effective to ask children to enact concepts or gestures in a more structural manner.

VI. Conclusions

In this paper, we present a large-scale study in which social robots try to teach preschool children words in a foreign language. The aims of the study were to investigate to what extent social robots can be effective when used in structured one-to-one tutoring sessions, whether robots would be more effective than a tablet application, and whether iconic gestures would be beneficial. The results demonstrate that robots can be effective tutors, but they are inconclusive about the added value compared to a tablet application and about the use of iconic gestures.

One of the main features of this experiment is the scale of the study and the fact that it is preregistered. While our large-scale study has not yielded the conclusions we have hoped for, this study is nevertheless extremely valuable in demonstrating the limitations and opportunities of using social robots as second language tutors in ways that would not have been feasible in smaller-scale studies. For example, the process of developing this experiment has taught us a lot about the issues involved in setting up such a large-scale experiment. Experiments which we believe are necessary to increase the credibility and acceptability of introducing social robots to address societal challenges, especially when it comes to health care and education.

Acknowledgments

To be further included in final version. We are also extremely grateful to all the schools, children and their parents who participated in this experiment.

References

[1] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science Robotics*, vol. 3, no. 21, p. eaat5954, 2018.

[2] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati, "The physical presence of a robot tutor increases cognitive learning gains," in *Proc of the 34th Annual Conf of the Cognitive Science Society*, 2012.

[3] I. Leite, M. McCoy, M. Lohani, D. Ullman, N. Salomons, C. Stokes, S. Rivers, and B. Scassellati, "Emotional Storytelling in the Classroom: Individual versus Group Interaction between Children and Robots." ACM Press, 2015, pp. 75–82.

[4] T. Belpaeme, P. Vogt, R. Van den Berghe, K. Bergmann, T. Göksun, M. De Haas, J. Kanero, J. Kennedy, A. C. Küntay, O. Oudgenoeg-Paz *et al.*, "Guidelines for designing social robots as second language tutors," *International Journal of Social Robotics*, pp. 1–17, 2018.

[5] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial," *J Hum Comp Interact*, vol. 19, no. 1, pp. 61–84, 2004.

[6] S. Lee, H. Noh, J. Lee, K. Lee, G. G. Lee, S. Sagong, and M. Kim, "On the effectiveness of robot-assisted language learning," *ReCALL*, vol. 23, no. 01, pp. 25–58, 2011.

[7] J. K. Westlund and C. Breazeal, "The Interplay of Robot Language Level with Children's Language Learning during Storytelling." ACM Press, 2015, pp. 65–66.

[8] J. Kennedy, P. Baxter, E. Senft, and T. Belpaeme, "Social Robot Tutoring for Child Second Language Learning," in *Proc of the 11th ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2016, pp. 67–74.

[9] M. Alemi, A. Meghdari, and M. Ghazisaedy, "The Impact of Social Robotics on L2 Learners' Anxiety and Attitude in English Vocabulary Acquisition," *Int J Social Robot*, pp. 1–13, 2015.

[10] K. Dautenhahn, "Human-robot interaction," *The Encyclopedia of Human-Computer Interaction, 2nd Ed.*, 2013.

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

[12] L. M. Marulis and S. B. Neuman, "The Effects of Vocabulary Intervention on Young Children's Word Learning: A Meta-Analysis," *Rev Educ Res*, vol. 80, no. 3, pp. 300–335, 2010.

[13] S. Lee, H. Noh, J. Lee, K. Lee, and G. G. Lee, "Cognitive effects of robot-assisted language learning on oral skills," in *INTERSPEECH 2010 Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*, 2010.

[14] F. Tanaka and S. Matsuzoe, "Children Teach a Care-Receiving Robot to Promote Their Learning: Field Experiments in a Classroom for Vocabulary Learning," *J Hum Robot Interact*, vol. 1, no. 1, pp. 78–95, 2012.

[15] 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.

[16] H. H. Clark, *Using Language*. Cambridge University Press, 1996.

[17] G. Pezzulo, L. W. Barsalou, A. Cangelosi, M. H. Fischer, K. McRae, and M. Spivey, "Computational grounded cognition: a new alliance between grounded cognition and computational modeling," *Frontiers in psychology*, vol. 3, p. 612, 2013.

[18] A. M. Glenberg, "Embodiment for education," *Handbook of cognitive science: An embodied approach*, pp. 355–372, 2008.

[19] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[20] M. Tomasselo and J. Todd, "Joint attention and lexical acquisition style," *First Lang*, vol. 4, pp. 197–212, 1983.

[21] M. Tellier, "The effect of gestures on second language memorisation by young children," *Gestures in Language Development*, vol. 8, no. 2, pp. 219–235, 2008.

[22] M. Macedonia and K. von Kriegstein, "Gestures enhance foreign language learning," *Biolinguistics*, vol. 6, no. 3-4, pp. 393–416, 2012.

[23] M. Macedonia, K. Bergmann, and F. Roithmayr, "Imitation of a pedagogical agents gestures enhances memory for words in second language," *Science Journal of Education*, vol. 2, no. 5, pp. 162–169, 2014.

[24] S. D. Kelly, T. McDevitt, and M. Esch, "Brief training with co-speech gesture lends a hand to word learning in a foreign language," *Language and Cognitive Processes*, vol. 24, no. 2, pp. 313–334, 2009.

[25] B. Irfan, J. Kennedy, S. Lemaignan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Social psychology and human-robot interaction: An uneasy marriage," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2018, pp. 13–20.

[26] 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 *Proc of the 12th ACM/IEEE Int Conf on Human-Robot Interaction*. ACM, 2017, pp. 82–90.

[27] J. Kanero, O. E. Demir-Lira, S. Koskulu, G. Oranç, I. Franko, A. C. Küntay, and T. Göksun, "How do robot gestures help second language learning?" in *Earli SIG 5 Abstract book*, 2018.

[28] L. M. Dunn, L. M. Dunn, and L. Schlichting, *Peabody picture vocabulary test-III-NL*. Amsterdam: Pearson, 2005.

[29] H. Mulder, H. Hoofs, J. Verhagen, I. van der Veen, and P. P. M. Leseman, "Psychometric properties and convergent and predictive validity of an executive function test battery for two-year-olds," *Frontiers in Psychology*, vol. 5, p. 733, 2014.

[30] S. Chiat, "Non-word repetition," in *Methods for assessing multilingual children: Disentangling bilingualism from language impairment*, . N. M. E. S. Armon-Lotem, J. de Jong, Ed. Bristol: Multilingualism Matters, 2015, pp. 227–250.

[31] J. Kory Westlund, L. Dickens, S. Jeong, P. Harris, D. DeSteno, and C. Breazeal, "A comparison of children learning new words from robots, tablets, & people," in *Proceedings of the 1st Int Conf on Social Robots in Therapy and Education*, 2015.

[32] J.-A. Mondria and B. Wiersma, "Receptive, productive, and receptive+ productive l2 vocabulary learning: What difference does it make," *Vocabulary in a second language: Selection, acquisition, and testing*, vol. 15, no. 1, pp. 79–100, 2004.

[33] T. Schodde, K. Bergmann, and S. Kopp, "Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making," in *Proc of the 12th ACM/IEEE Int Conf on Human-Robot Interaction*. ACM, 2017.

[34] 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.

[35] J. Kennedy, P. Baxter, and T. Belpaeme, "The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning," in *Proceedings of the tenth annual ACM/IEEE International conference on Human-Robot Interaction*. ACM, 2015, pp. 67–74.