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

Paul Vogt*, Rianne van den Berghe†, Mirjam de Haas*, Laura Hoffman‡, Junko Kanero§, Ezgi Mamus§, Jean-Marc Montanier¶, Cansu Oranç§, Ora Oudgenoeg-Paz†, Daniel Hernández García‖, 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†, Amit Kumar Pandey¶

*Tilburg University, Tilburg, the Netherlands †Utrecht University, Utrecht, the Netherlands

‡Bielefeld University, Bielefeld, Germany §Koç University, Istanbul, Turkey

¶SoftBank Robotics, Paris, France ‖University of Plymouth, Plymouth, United Kingdom

**Ghent University, Ghent, Belgium

Email: *p.a.vogt@uvt.nl

***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 (supported by a tablet), 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 accompanied by a tablet versus learning from a tablet application alone. For reasons of transparency, the research questions, hypotheses and methods were preregistered. With a sample size of 194 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 beneficial effect of a robot's iconic gestures on learning gains.**

***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], [2]. Next to science, technology, engineering and mathematics (STEM) topics, (second) language tutoring is seen as an area for which robots can offer effective educational support [3]–[6]. Robots, by being physically present in the students' environment, are able to provide effective one-on-one tutoring [7], which can result in significantly higher learning gains than group-based education [8]. This is facilitated by the robot's ability to exhibit socially supportive behaviour [9]. However, it is still unclear to what extent robots can be effective language tutors, and how to best design effective robot language tutors. We believe that one reason for this is that many recent studies are statistically underpowered and/or often glean results from only a single interaction session (e.g., [10], [11]).

The reason for this is that the development and execution of human-robot interaction (HRI) experiments is time-consuming and costly, especially for long-term interaction studies [12]. 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" [13]. 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 [14] for an overview), and the learner tends to lose interest in the robot. Long-term studies are particularly critical for educational robots, because learning a particular skill, such as speaking and understanding a second language (L2), requires repetition and time [15].

Few studies have investigated the effect of robots in multiple interactions on language learning [4], [6], [16], [17], with mixed results. For instance, Kanda et al. [4] did not observe a clear learning effect in their two-week field trial where children were taught English, 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 [4]. On the other hand, studies by Movellan et al. [17] and Gordon et al. [16] have demonstrated that children can learn a limited number of new words from a robot over the course of multiple interactions.

Many of these (long-term) HRI studies, however, are relatively exploratory in nature due to small sample sizes and/or a limited number of experimental conditions. To study, 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, ideally over a longer period of time and over the course of multiple interactions.

This brings us to the following question: to what extent can social robots be effective in L2 tutoring on the long term? Moreover, are they more effective than other digital (screen-based) tutors; if so, why? A good argument for why

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No. 688014.

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 [18]. Compared to other screen-based technologies, the interactions with a physical robot provide such grounding and are situated in a three-dimensional, tangible world [19]. The physicality of the interaction allows for a true implementation of the embodied cognition paradigm, which states that our cognitive processes, such as language comprehension and scientific thinking, are supported by our bodily experiences (e.g., perceiving and acting in the real world) [20].

One of the features in which the physicality of the interaction can manifest itself is by having robots interact multimodally, by communicating both verbally and non-verbally. In gesture research, one often distinguishes deictic gestures (such as pointing) from iconic gestures (where the shape of the gesture has some physical similarity to its referent) [21]. 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 [22]. Iconic gestures produced by tutors can also have a positive effect on vocabulary learning in children [23] and in adults [24], [25], also when the gestures are produced by robots [11]. 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 [26] 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 [20].

In the current study, which is part of the L2TOR project[1], we investigate the effect that robots—either using iconic and deictic gestures or only deictic gestures—may have on teaching 5- to 6-year-old children basic vocabulary from a foreign language in a longitudinal study over seven sessions. Moreover, the effect of interacting with a robot tutor supported by a tablet game is compared to interacting with a tablet game without a robot. In contrast to many other previous studies, the study is statistically well-powered with a sample size of 194 children. The experiment has four conditions:

1) *Robot with iconic gestures + tablet* where the robot supports tutoring using iconic and deictic gestures, and with interactions mediated by a tablet game.
2) *Robot without iconic gestures + tablet* 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 its speech output routed through the tablet's speakers, and with interactions mediated by a tablet game.
4) *Control* condition where children danced with the robot but were not exposed to the educational material.

The control group is included as a "non-treatment" condition, receiving an activity related to what is occurring in the experimental conditions (multiple one-on-one interactions with a robot), but not related to the goal of the experiment (teaching English words). Furthermore, the addition of this condition controls for the possibility that children learn English words without directly being taught (e.g., from the tests that were administered). Therefore, any difference found between the control and experimental groups shows the effect of the intervention.

In this paper, we investigate the effect that the different conditions have on learning gains. Based on predictions both from the aforementioned literature and earlier studies with robot tutors, we formulate the following hypotheses:

H1: The robot will be effective in teaching children L2 target words: children will learn more words from a robot (**H1a**) and will remember them better (**H1b**) than children who participate in a control condition — comparison between the results of conditions (1) and (2) with condition (4).

H2: Children will learn more words (**H2a**), and will remember them better (**H2b**) when learning from a robot and a tablet than from a tablet only — comparison between the results of conditions (1) and (2) with condition (3).

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 — comparison between the results of condition (1) with condition (2).

The research questions, hypotheses, and methods have been preregistered at AsPredicted[2]. 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 [27].

In the remainder of this paper, we first outline the lesson plan and the basic interactions we designed between the child, 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 a NAO robot as a (nearly) autonomous tutor. All lessons involved one-on-one interactions between robot and child. Interactions were mediated through a game played on a Microsoft Surface touch-screen tablet computer, which provided visual context and recorded touch-based input from the child. This tablet interface was included, because the implementation of fully

[1]http://www.l2tor.eu

[2]https://aspredicted.org/6k93k.pdf

Fig. 1. The basic setup for all lessons.

autonomous social robot behaviour in a complex and dynamic environment is challenging [28], and specifically because there is no reliably performing automatic speech recognition for children's speech [29]. 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 was placed in a 90-degree angle towards the child, so that the robot and the child had a similar frame-of-reference. A video camera placed on a tripod facing the child was used to record the interactions. A second camera was placed from the side to get a more complete overview of the interactions.

A. Target words

English target words were selected from 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 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.

A total of 34 target words were selected. Selection was based on school curricula, child-language corpora, and age-of-acquisition lists containing the average age at which a particular word in a language is acquired. 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. We confirmed that children indeed knew all 34 words in Dutch by pilot testing the materials for the pre-test in Dutch with 15 native speaking 5- to 6-year-olds.

The 34 target words were introduced to the children in six lessons, each including five or six words, and were recapped in a seventh 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

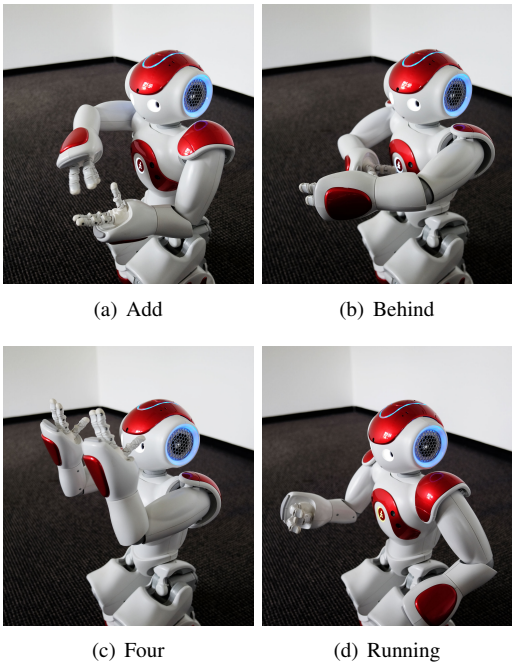
Each of the six content lessons consisted of three phases. The first phase was a brief introduction where the robot would greet the child by name, and present the new virtual environment (e.g., playground) that set the context of the lesson. The second phase was a word modelling phase where the target words of the current lesson were named for the first time, mapping the concepts in English to their Dutch equivalents together with a visual example 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). For example, a group of two elephants would appear on the screen, which the robot then narrated in Dutch ("Look, elephants!"), before asking the child to touch the elephants to find out the English word for the concept (*two*). Upon touching the objects, the English word for *two* was then introduced by the tablet through a pre-recorded audio clip of a female native English speaker pronouncing the word *two*. The robot would then repeat the word and ask the child to repeat the target word as well. 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 reliably.

After a target word was introduced, the robot and child would engage in certain tasks that revolved around the target word. For instance, the child was asked to 'add' 'one', 'two' or 'three' animals in a cage. 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. If children did not perform a certain task, the robot asked the child once more. After two such reminders, or if the child performed the task incorrectly twice in a row, the robot offered to help. In case of a manipulation on the tablet (e.g., touching or moving an object), the robot 'magically' demonstrated 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. If the task required the child to repeat a word or phrase, the robot counted down from three to one and said the word or phrase together with the child. The lesson then proceeded irrespective of the child's response.

Once all target words were modelled, each lesson ended 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 or actions from that specific lesson, and the child was asked to tap on the relevant picture or animation. During these tests, the robot did not provide any feedback, nor gestures, to help children. The results of these tests are beyond the scope of this paper.

The seventh session was a recap lesson, where children created a picture book on the tablet. They were presented with, one by one, the scenes of the six content lessons, and 'stickers' with the target objects of these lessons. The children placed these 'stickers' on the scenes, while the robot talked about 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 the learner's perspective. (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. Videos of these examples are available at https://youtu.be/Ebz2fLKVFsg

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 + tablet: In this condition, the robot produced 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 [30].

2) Robot without iconic gestures + tablet: Here, the robot did 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 produced a deictic gesture. Part of these deictic gestures, on average eight per lesson, were planned at specific times in order to draw the child's attention to an activity happening on the tablet. The other deictic gestures occurred when the child did not respond to an instruction to manipulate something on the tablet, where the robot performed the aforementioned 'magical' demonstration of how to execute the task. The total number of deictic gestures therefore varied based on the amount of help a child needed from the robot.

3) Tablet-only: In this condition, the robot was hidden from the child's view. The robot's voice was directed to come from the tablet's speakers and the information displayed on the tablet was exactly the same as in the two robot conditions. To compensate these children for not seeing the robot during the experiment, we organised a group session with the robot, similar to the introduction (see next section), after the immediate post-test was administered.

4) Control: Children in the control condition were not exposed to any of the lesson content, but would instead dance to a popular Dutch children's song once per week—a different song each time—over the course of three weeks. The choice for three weeks of one-on-one sessions was made in order to align with the experimental conditions, where children would also receive all six content lessons over the course of three weeks.

III. METHODS

A. Participants

A total of 208 children, all native speakers of Dutch, were recruited from nine different primary schools in the Netherlands. The average age was 5 years and 8 months ($SD = 5$ months). To ensure that their prior knowledge of English was not consequential, children could only participate if they would not exceed a score of 17 out of 34 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, nine children dropped out for various reasons, such as fussing and shyness. Data of an additional two children were excluded as they missed one lesson or had received one lesson twice due to technical issues. The resulting sample included 194 children. Table II shows how the final set of participants are distributed 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 then asked to state what it meant in Dutch. 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 children's ability to learn language:

- Dutch vocabulary knowledge (Peabody Picture Vocabulary Test) [31],
- selective attention (visual search task) [32], and
- phonological memory (non-word repetition task) [33].

2) Post-tests: We conducted two post-tests (one immediate post-test, administered maximally two days after the final lesson, and one delayed post-test, which took place between two and four (M = 2 weeks and 5 days, SD = 2.70 days) weeks after the seventh lesson). Both post-tests contained three parts:

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

TABLE II
OVERVIEW OF THE PARTICIPANTS IN THE EXPERIMENT.

Condition	N	Gender N_b/N_g	Avg Age + SD (Y;M)	(M)
Iconic gestures	54	31/23	5;8	5
No iconic gestures	54	28/26	5;8	5
Tablet-only	54	24/30	5;9	5
Control	32	14/18	5;7	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 short video-clips simultaneously and asked to choose the picture or video corresponding to the target word. Each target word was tested three times, using a different picture or video-clip and using different distractor images. This is a standard way to test word knowledge in language learning studies to reduce the bias that may result from guessing. However, since doing this for all 34 target words would be too taxing for the child, 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.) and words from all lessons. 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 not discussed in the current paper.

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. This framing as a peer was done because previous research has shown that children perform better and appear to show greater ability to focus their attention when the robot behaves like a peer rather than a tutor [34]. 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 an orifice that resembles a mouth but that this 'mouth' does not move when it speaks, and that although it has big 'ears', they should speak loud and clearly to its face when they want the robot to understand them. 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 were rewarded with stickers 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 guided from their classroom to a room dedicated to the experiment. The child was instructed to sit in front of the tablet and in a 90-degree angle with the robot (see Fig. 1) after which the researcher would start the lesson. During the first part of the lessons, the researcher would help children if needed by encouraging them to touch the display or telling them that it is their turn to answer the robot. Otherwise, the researcher would sit somewhere behind the child and operate

the wizard to ensure the interaction would proceed as planned when the child responded verbally to the robot's request. If the child had to go to the bathroom or in the event of a system crash (which happened infrequently), the lesson was paused and would continue after the child had returned or the system was rebooted. At the end of each lesson, the child was (virtually) rewarded with a star and brought back to the classroom. The duration of the experimental sessions varied per lesson and per condition, taking on average 17 minutes and 16 seconds (SD = 3 minutes, 47 seconds); with lesson 7 (the recap lesson) taking longest and lesson 1 being the shortest. Lessons in the iconic gesture condition took the longest (M = 20 minutes and 59 seconds, SD = 1 minute and 49 seconds), followed by the tablet-only condition (M = 15 minutes and 37 seconds, SD = 1 minute and 8 seconds) and the no iconic gestures condition (M = 15 minutes and 6 seconds, SD = 1 minute and 13 seconds). The sessions of the control condition were significantly shorter and only took about five minutes per session.

After all seven lessons were completed, the two post-tests were administered by a trained researcher. As with the pre-tests, the post-tests were administered in one-on-one sessions using paper score sheets. The immediate post-test took about 40 minutes, while the delayed post-test took 30 minutes. The difference between the two tests was that the questionnaire regarding the child's perception of the robot was only included in the immediate post-test.

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 word-knowledge 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; $t(193) = 16.45; p < .001$). All other translation tasks from the two post-tests also differ significantly from zero (all p values $< .001$). 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 translation scores from English to Dutch measured in the immediate post-test are higher than those measured in the pre-test for all experimental conditions (all p values $< .001$) 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 (all p values $< .001$).

To test our hypotheses, we performed a 2 (post-test moments) $\times$ 4 (conditions) repeated measures analysis using a doubly multivariate design, applied simultaneously to both the mean of the translation tasks and the comprehension task. The findings showed a main effect of condition ($F(6, 378) = 3.34, p = .003, \eta_p^2 = .05$). Bonferroni-corrected post-hoc tests showed that children in the experimental conditions scored higher than children in the control condition on all tasks (all p values $< .01$), but there were no significant differences between the experimental conditions (all p values $> .10$). Also, a main effect of time revealed that scores of the delayed post-test were significantly higher than those of the immediate post-test ($F(2, 190) = 5.99, p = .003, \eta_p^2 = .06$), suggesting that newly learned words need time to become consolidated.

TABLE III
The main test results.

Condition / Test	Pre-test	Imm. post-test	Delayed post-test
Iconic gestures			
Trans(En-Du)	3.31 (3.09)	7.41 (5.17)	8.10 (5.06)
Trans(Du-En)		6.00 (4.23)	6.45 (4.62)
Comprehension		29.47 (5.85)	30.43 (6.22)
No iconic gestures			
Trans(En-Du)	3.47 (3.19)	7.69 (4.92)	7.88 (4.79)
Trans(Du-En)		6.43 (4.20)	6.43 (4.65)
Comprehension		29.39 (6.08)	29.75 (6.44)
Tablet-only			
Trans(En-Du)	4.04 (2.76)	7.96 (4.63)	8.63 (4.62)
Trans(Du-En)		6.57 (4.01)	6.67 (4.20)
Comprehension		29.73 (6.27)	30.25 (6.58)
Control			
Trans(En-Du)	2.48 (2.25)	3.48 (2.75)	3.97 (2.82)
Trans(Du-En)		3.07 (2.27)	3.52 (2.17)
Comprehension		24.31 (6.25)	25.62 (5.34)

Note: All scores indicate the average number of words correctly translated or comprehended (standard deviation within brackets). Minimum scores are 0, maximum scores are 34 for translation and 54 for comprehension. For comprehension, chance level is 18.

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 two multiple regression analyses with the mean score on the translation tasks and the comprehension task 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 .16 and .17, all p values $< .05$). 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 delayed post-test as dependent variables, we observed the same trend for the translation tasks. However, there was no effect of vocabulary on scores of the comprehension task.

V. Discussion

In this paper, we presented a large-scale evaluation study that was conducted in order to investigate to what extent social robots can contribute to L2 tutoring in early childhood. We compared two different robot conditions—one with and one without iconic gesturing—with a control group in which children did not receive any language tutoring. Furthermore, we investigated the added value of the robot's physical presence by comparing the two robot conditions—both including a tablet computer for display and interaction purposes—with

a condition in which children received the same input via the tablet and in which no robot was present.

This study is unique in three 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 seven consecutive lessons in which they were taught a total of 34 English words. (2) This study was statistically well-powered with a total of 194 children participating in one of four conditions. (3) The experiment's research questions, methods and hypotheses were preregistered online to ensure transparency about the way our study was planned, and the way data were collected and analysed.

To summarise the main findings, we found evidence to support hypothesis H1 that children learn L2 target words from a tutoring robot with a tablet, and that they can remember them better than children who participate in a control condition where they were not exposed to the lessons. However, we did not find evidence to support hypothesis H2 that children will learn more words and remember them better when learning from a robot and a tablet than from a tablet alone. In fact, the results indicate that children learn equally well from the robot and the tablet as from just the tablet. The combination of these findings indicates that the success of the tutoring sessions (H1) cannot be attributed solely to the interactions with the robot.

Finally, we also did 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 [11], [23], [24], the present study does not confirm this. In the remainder of this section, we will elaborate on these findings.

A. Social robots versus 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 + tablet conditions as in the tablet-only condition. To evaluate these findings, it is important to understand the similarities between the conditions. All interactions in the two robot conditions were mediated by the tablet, which displayed the learning context and recorded the child's input and responses to the system. Essentially, the children played educational games on the tablet in all conditions. In the two robot conditions, the robot provided 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 because the robot's voice was directed through the tablet's speakers, but the non-verbal support was not provided and there was no agent visually present. It was a conscious, methodological design choice to keep the interactions as similar as possible between robot + tablet and tablet-only conditions, in order to measure the effect of the robot's physical presence and its ability to provide non-verbal support without introducing any confounding factors.

The decision to have the majority of these interactions take place within the context of an educational game on a tablet was made in order to achieve a fully autonomous system, thereby circumventing technical challenges with the robot's sensing abilities such as automatic speech recognition for children [29] and object tracking in an unconstrained environment [35]. However, the substantial role of the tablet may have somewhat limited the importance of the interaction between child and robot. In the tablet condition, children could focus their attention solely on the tablet game, while attention had to be divided between the two devices in the robot + tablet conditions. This may have negatively influenced the potential contribution of the robot's non-verbal support to L2 learning, affecting the successful integration of embodied cognition as a result.

However, it is plausible that children found interacting with the robot more engaging than with a tablet only, even though it did not systematically influence their learning. To further investigate this, we are currently analysing children's engagement in two dimensions: engagement with the learning tasks, and their social engagement with the robot, based on video recordings of the study, to investigate how these two types of engagement varied over the different conditions.

B. Iconic gesturing

Given that research has shown that iconic gestures can help people learn vocabulary in an L2 [23], [24], even when supplied by a social robot [11], we expected to observe a similar effect in this experiment. However, our hypothesis on this issue was not supported. In addition to the previously mentioned substantial role of the tablet game in the interaction, the design of the gestures could have negatively affected their contribution to the learning process. Initially proposed by adults, the gestures were implemented on the robot after which their clarity was rated by other adults [30]. However, some gestures had to be altered because in the original versions the robot was standing up, whereas in the current study the robot was in a crouched position. Moreover, the positioning of the robot at a 90-degree angle may have negatively affected the clarity of the gestures. Anecdotally, we observed a child trying to mirror the robot's gesture for the word *two* by actually showing three fingers instead of two. To verify if the angle at which the gestures were viewed has affected their comprehensibility, we are currently conducting a survey in which adult participants are shown recordings of all gestures at the same 90-degree angle used in the study (see Figure 2), and are asked to guess which concept is being portrayed and rate the clarity of the gesture.

Another reason why iconic gestures may not have yielded the expected effect is that they were shown every time the robot mentioned one of the target words, which was at least ten times per target word per lesson. 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. [36]). 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 or when additional support is needed, for example when a child is having difficulty with a particular concept. This ties into our ambition to include a larger degree of adaptivity and personalization in future work, where certain parts of the tutoring interaction (e.g., the use of gestures, pacing or difficulty of the tasks) are changed based on the child's performance or affective state [11], [37].

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 [20]. In our experiment, children were only in some sessions asked to enact a certain concept (e.g., running). Therefore, we are analysing to what extent children re-enacted the gestures without being prompted to, 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.

C. On the experimental design

The expectation resulting from our literature review was that children could learn L2 words from a social robot over multiple lessons [5], [16], [38]. By comparing the results between the experimental robot conditions and the control condition we indeed demonstrated that our implementation of a tutoring system was effective at teaching the children new vocabulary, and that this was not caused by external factors (e.g., activities occurring at the school or at home).

However, children in the control condition did score higher on the two post-tests than on the pre-test in the English to Dutch translation task, and they also scored significantly higher on the delayed post-tests than on the immediate post-tests. This demonstrates that these children, despite not having received any lessons from the robot, did also learn English vocabulary. In other words, although the experimental intervention was responsible for a significant (though relatively modest) learning effect, external factors seemed to have caused some learning to occur as well. This could have originated from participating in the tests, from interacting with other children who were in the experimental (i.e., non-control) conditions or from being exposed to English vocabulary elsewhere; after all, most children who participated in the study already knew some of the English target words prior to the experiment.

We are confident that the translation and comprehension tasks that were used in this study provided a reliable measurement of a child's knowledge of the English target words. The scores of the English to Dutch translation tasks were on average around 8 out of 34 in the two post-tests of the experimental conditions. Although this may seem low, these findings are consistent with those in earlier studies on L2 word learning demonstrating low scores in translation tasks [39]. Translating words from Dutch to English seems even more difficult, yielding an average score of around 6.5 out of 34 in all experimental conditions. Comprehension scores were considerably higher, which is not surprising, given that comprehension tasks are generally easier: the learner only has to recognize the target word from a small set of pictures or videos, instead of retrieving and producing the word themselves without context. In our study, children performed significantly better than chance on this task, and children in the experimental conditions performed significantly better than children in the control condition.

VI. Conclusions

In this paper, we presented a large-scale study in which a social robot was used 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-on-one tutoring sessions, whether robots would be more effective than a tablet application, and whether iconic gestures would be beneficial. The results demonstrated that the tutoring interaction, consisting of a robot and a tablet game, was effective, but they did not show an added value of the robot compared to using only a tablet application, nor of the use of iconic gestures as they were implemented in this study. Several design choices were made during development of the experiment (e.g., regarding the tablet interactions and iconic gestures), which have been documented in this paper. In follow-up studies, we intend to re-evaluate these decisions and investigate how they may have affected the presented findings.

Arguably, the main contributions of the research presented in this paper are the scale (i.e., sample size and long-term nature) of the experiment and the fact that the study was preregistered. While our large-scale study has not yielded the conclusions we predicted, it 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. There is a clear need of more large-scale experiments in order to increase the credibility, acceptability and effectiveness of introducing social robots to address societal challenges, especially when it comes to healthcare and education.

Acknowledgments

We are very thankful to Laurette Gerts, Annabella Hermans, Esmee Kramer, Madée Kruijt, Marije Merckens, David Mogendorff, Sam Muntjewerf, Reinjet Oostdijk, Laura Pijpers, Chani Savelberg, Robin Sonders, Sirkka van Straalen, Sabine Verdult, Esmee Verheem, Pieter Wolfert, Hugo Zijlstra and Michelle Zomers for their help in collecting and processing the data of this experiment. A special thanks to Chrissy Cook for lending us her lovely voice for the audio recordings. We are also extremely grateful to all the schools, children, and parents of the children that 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," *Educational Psychologist*, vol. 46, no. 4, pp. 197–221, 2011.

[9] M. Saerbeck, T. Schut, C. Bartneck, and M. D. Janse, "Expressive Robots in Education: Varying the Degree of Social Supportive Behavior of a Robotic Tutor," in *Proc of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 1613–1622.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Affective Personalization of a Social Robot Tutor for Childrens Second Language Skills," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[17] J. R. Movellan, M. Eckhardt, M. Virnes, and A. Rodriguez, "Sociable robot improves toddler vocabulary skills," in *Proc of the 4th ACM/IEEE Int Conf on Human-Robot Interaction*, 2009, pp. 307–308.

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

[19] 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.

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

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

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

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.