Second Language Tutoring using Social Robots

Project No. 688014

L2TOR

Second Language Tutoring using Social Robots

Grant Agreement Type: Collaborative Project
Grant Agreement Number: 688014

D7.2 Evaluation Report Space Domain

Due Date: **30/09/2018**
Submission Date: **12/11/2018**

Start date of project: **01/01/2016** Duration: **36 months**

Organisation name of lead contractor for this deliverable: **UU**

Responsible Person: **Ora Oudgenoeg-Paz** Revision: **1.0**

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

Contents

Executive Summary 3
Principal Contributors 4
Revision History 5
1 Introduction 6
2 Descriptive analyses large scale evaluation study 8
3 Analyses of pre-registered research questions large scale evaluation study 10
4 In-depth analyses of learning gains 15
5 Immediate learning gains 21
6 Perception of the robot and learning gains 26
7 Discussion and conclusion 28
Reference List 32
Appendix I: Vogt et al., submitted 35
Appendix II: Results of in-depth analyses: Translation task L2>L1 45
Appendix III: Results of in-depth analyses: Translation task L1>L2 52
Appendix IV: Results of in-depth analyses: Comprehension Task 59
Appendix V: Results of in-depth analyses: Test at the end of each lesson 71
Appendix VI: Van den Berghe et al., submitted 73

Executive Summary

This deliverable includes first results of the large-scale evaluation study, conducted from February till June 2018. In our study children received 7 lessons where 34 English words were taught. Children were divided in 4 conditions : (1) being taught the words with a tablet and a robot that used both deictic and iconic gestures; (2) being taught the words with a tablet and a robot that used only deictic gestures; (3) being taught the words with a tablet only; (4) control group doing dance activities with the robot but not being taught any words. In this deliverable we present the first results of this study including analyses of the main hypotheses as described in the preregistration, in-depth analyses of learning gains for individual words and children's perception of the robot and its relation to learning gains.

Results show that our training was effective, meaning that children in all three experimental conditions did learn English words. Learning gains were, however, smaller than in traditional vocabulary training programmes. Moreover, we found no differences between the three experimental conditions. Meaning that, in this experiment, the use of a robot had no added value to just using a tablet and that the use of iconic gestures had no added value to using just deictic gestures. Given that working ASR for children and object recognition are not yet feasible in the current state of technology, we resorted to working with a tablet to mediate the interaction between the robot and the child. Moreover, these limitations, in combination with the controlled experimental nature of the study meant that our training was static. This entails that all children received almost identical lessons, varying only in the amount of feedback. A more adaptive system that can adjust teaching strategies and level of information taught (like what a human tutor does), will likely be more effective. Moreover, the use of a tablet, which could not be avoided, might impede a more natural interaction between the child and the robot. Such a natural interaction might do more justice to the potential of the robot as an embodied agent. The use of a tablet might also explain the lack of effect of the use of iconic gestures, as children primarily interacted with the tablet. In the last section of this deliverable we discuss the meaning of the results and suggest directions for further analyses and future research.

Principal Contributors

UU: Esmee Kramer, Ora Oudgenoeg-Paz, Rianne van den Berghe, Paul Leseman, Josje Verhagen, Michelle Zomers
TIU: Mirjam de Haas, Emiel Krahmer, Paul Vogt, Bram Willemsen, Jan de Wit
UNBI: Thorsten Schodde, Larua Hoffmann, Stefan Kopp, Kirsten Bergmann
ALD: Jean-Marc Montanier, Amit Kumar Pandey
KOC: Junko Kanero, Canso Ornaç, Tilbe Göksun, Aylin Kuntay
PLYM: Fotios Papadopoulos, Christopher Wallbridge, Tony Belpaeme

Revision History

Version 1.0 (OO 12-11-2018)
This is the first version.

1 Introduction

As a result of the revised objectives of the L2TOR study as approved by the project officer and reviewers, the content of the current deliverable deviates from the title. Instead of a final evaluation of the space domain, this deliverable includes first results of the large-scale evaluation study, conducted from February till June 2018. The design and rationale of this study were extensively described in D7.1 and now all data have been collected. Therefore, we provide only a summary of the study below.

While social robots hold the promise to be effective in tutoring a second language to preschool children, their effectivity has not yet convincingly been demonstrated in empirical studies, and it is not clear which characteristics are important for such robot tutors. One reason for this lack in knowledge is the fact that current studies often use small samples preventing drawing firm conclusions, and often include only a single interaction between a child and the robot. Moreover, most studies use no comparison to either a control group or a group of children studying with other, more traditional, digital technologies, such as tablets. To address these issues, we conducted a large-scale study in which Dutch preschool children were taught an L2 (English) in multiple one-on-one tutoring sessions with a social robot.

In our study children received 7 lessons (6 “core” lessons where target words in English were taught and 1 recap session) containing 34 English target words in total. Children were divided in 4 conditions pseudo-randomly (taking gender into account): (1) being taught the words with a tablet and a robot that used both deictic and iconic gestures; (2) being taught the words with a tablet and a robot that used only deictic gestures; (3) being taught the words with a tablet only; (4) control group doing dance activities with the robot but not being taught any words.

About one week before the start of the lessons, children received a group introduction where they saw the robot, were taught how to interact with it (e.g., they were taught that it cannot hear very well and that they had to speak loudly while facing the robot). They also engaged in a game and a dance with the robot. They were told that the robot was called Robin the robot and it was framed as a peer who would learn English with them.

After the introduction, but before the start of the lessons, a pretest was conducted to test children’s knowledge of the target words in English. Children heard each of the English target words, and were asked what it meant in Dutch (“Wat betekent [target word] in het Nederlands?” / What does [target word] mean in Dutch?). The English words were pre-recorded by a native English speaker female. The test was administered using a laptop computer. Additionally, several cognitive skills known to be associated with language learning skills were measured: Dutch receptive vocabulary (Werker & Yeung, 2005), selective attention (Ellis, 2006), and phonological memory (Gathercole, 2006; Masoura & Gathercole, 2005; Verhagen & Leseman, 2016). Additionally, we tested children’s perception of the robot with a short questionnaire consisting of 19 questions. Children were asked whether they thought certain human characteristics applied to the robot, such as being able to see, think, feel, grow, et cetera (based on Jipson & Gelman, 2007).

At the end of each of the 6 lessons in which target words were taught, children did a short task aimed at measuring their knowledge of the target words taught in the lessons. In this task, children were presented with 3 images on the tablet screen and asked to click on the image that corresponded to the target word they heard. The (voice of the) robot did not provide help or feedback during this test. Each target word was tested twice using two different trials using different sets of images. The animations used for the items were similar to the animations used during the lessons.

After all 7 lessons were completed, we conducted two post-tests. The first post-test was administered maximally 2 days after the last lesson and a second post-test took place between 2 and 5 weeks after the last

lesson. The aim of the second post-test was to assess retention of the learned target words. In both post-tests we used a translation task from English to Dutch and from Dutch to English as well as a picture selection task used to measure receptive knowledge of English target words. The two translation tasks followed the same structure as the task of the pre-test and contained all 34 target words. In the comprehension (i.e., picture selection) test children saw three photos or films simultaneously and were asked to choose which image matched a (target) word. This was done by asking the question: 'where do you see [target word]? For prepositions a reference object was added to the question (e.g., 'where do you see [target preposition] the tree'). The carrying sentence was in Dutch and only the target words were in English (i.e., 'waar zie je [target word]?' / 'waar zie je [target preposition] de boom?'). The question were pre-recorded by a female bilingual native speaker of Dutch and English. Each target word was tested using three different trials containing different images for the target word and the distractors. Given the design of this task a test containing all 34 words would have been too long for such young children. Therefore only a selection of 18 target words was included. Words were chosen pseudo-randomly, by making sure that all word categories included in the lessons (i.e., verbs, prepositions, measurement words and so forth) were also represented in the test. During the first post-test the perception of the robot was measured again, using the questionnaire described above.

In this deliverable we present the first results of this large-scale evaluation study. First, in section 2 we describe the properties of the sample. Second, in section 3, we report the results of the analyses of the main hypotheses as described in the pre-registration (https://aspredicted.org/6k93k.pdf ; see also D7.1):

1. The robot will be effective at teaching L2 target words: children will learn words from a robot and will remember them better than children who participate in a no treatment (control) condition.
2. Children will learn more words and will remember them better when learning from a robot than from only a tablet.
3. Children will learn more words and will remember them better when learning from a robot that produces iconic gestures than from a robot that does not produce such gestures.

A report of these analyses has also been submitted to the International Conference on Human-Robot Interaction (HRI; Vogt et al., submitted) and is attached to this deliverable in Appendix I.

In section 4, we report the results of in-depth descriptive analyses at the word level. This analysis enabled us to compare the learning gains between target words. In addition, words were clustered together in word categories such as movement verbs, measurement words, prepositions, and count words, to examine differences in learning gains between categories.

In section 5 we report on an analysis of so called 'immediate learning gains'. These are the results of tests conducted at the end of every lesson, measuring knowledge of the words taught in these lessons. This analysis also enabled us to test whether learning gains are similar between lessons or not. Discussions in the literature about the novelty effect suggest that as children get accustomed to the robot and novelty wears off, learning gains might decrease (Kanda, Hirano, Eaton, & Ishiguro, 2004; Leite, Martinho, & Paiva, 2013).

Finally, in section 6 a study is presented on children's perception of the robot and its relation with learning gains. We examined whether children perceive the robot more as human or as a machine, does this perception change after having followed the lessons, and does this perception or the change in perception relate to learning gains. There is some evidence that people who perceive the robot more as a human tend to collaborate better with the robot (e.g., Duffy, 2003). Therefore, perception might influence learning gains. These results have been published in a paper submitted to HRI (van den Berghe, de Haas, et al., submitted; included in Appendix VI).

We conclude this deliverable with a general discussion session in section 7 in which we discuss the meaning of our findings thus far and suggest some possible explanations for the findings and directions for further analyses of the data.

2 Descriptive analyses large scale evaluation study

We recruited in total 208 children (50.5% girls) from 9 different primary schools in the north, middle, and south of the Netherlands. Children were all native speakers of Dutch and were on average 5 years and 8 months old (*SD* = 5 months) at the start of the lessons. Before the lesson series started, we tested children's knowledge of the 34 target English words that were included in the lessons series. Children who knew more than 17 (50%) of the words were excluded from the study. These children (*n* = 3) did a short dance with the robot to avoid them being disappointed that they do not get to play with the robot. During the lessons series, 9 children stopped participating for various reasons such as lack of compliance and shyness. Data of 2 additional children were excluded as they missed one lesson and received another lessons twice due to technical issues. The final sample therefore included 194 children. All parents or legal guardians signed informed consent and the children were rewarded a small gift after the last post-test to thank them for participation. Children who dropped out or were excluded also received the gift.

The children were pseudo-randomly divided in the four conditions, taking into account the gender distribution and the smaller number of children in the control group (group size was based on power analyses, see deliverable 7.1). The main characteristics of the participants in each condition are presented in Table 1.

Table 1
Characteristics of the participants

Condition	N	Mean age (Y;M) / SD (M)	Percentage of girls	PPVT(SD)	NWR(SD)
Robot using iconic gestures	54	5;8 / 5	43%	108.13(12.54)	10.08(2.97)
Robot not using iconic gestures	54	5;8 / 5	48%	108.67(11.83)	11.33(2.86)
Tablet only	54	5;9 / 5	55%	105.21(12.28)	11.08(2.15)
Control	32	5;7 /5	56%	108.88(13.96)	10.19(3.22)

Note. PPVT=Peabody Picture Vocabulary Test, results of a standard test measuring Dutch vocabulary (in the population *M* = 100, *SD* = 15). NWR=Nonword Repetition, a test measuring phonological memory (range of scores 0-16).

MANOVA and Chi square tests showed that, prior to the lesson series, children in the four conditions did not vary in age, gender, level of Dutch vocabulary, phonological memory and level of knowledge of the target words in English. Cronbach's alpha was also tested for the translation and comprehension tasks used in the pre- and post-tests. All values were above .85 with the majority being above .95. This indicates that reliability of the tasks is excellent. Table 2 shows the correlations between all the scores. The correlations are generally large. The correlations across measurement points suggest good test-retest stability. That is, children who knew more words during the pre-test also learned more during the training and children who

showed a higher level of knowledge at the first post-test also knew more at the delayed post-test. The correlations within a measurement point show that the two translation tasks correlate strongly, suggesting that they measured almost exactly the same skill. The correlations between the translation and comprehension tasks are generally medium suggesting that these tasks measure highly related, yet distinctive skills.

Table 2

Correlations between tasks used during pre- and post-tests

	1	2	3	4	5	6
1. Translation L2>L1 pretest						
2. Translation L2>L1 post-test 1	.74***					
3. Translation L1>L2 post-test 1	.66***	.85***				
4. Comprehension post-test 1	.61***	.75***	.67***			
5. Translation L2>L1 post-test 2	.74***	.91***	.82***	.75***		
6. Translation L1>L2 post-test 2	.67***	.83***	.88***	.70***	.85***	
7. Comprehension post-test 2	.60***	.69***	.65***	.70***	.74***	.70***

Note. ***$p < .001$

3 Analyses of pre-registered research questions large scale evaluation study

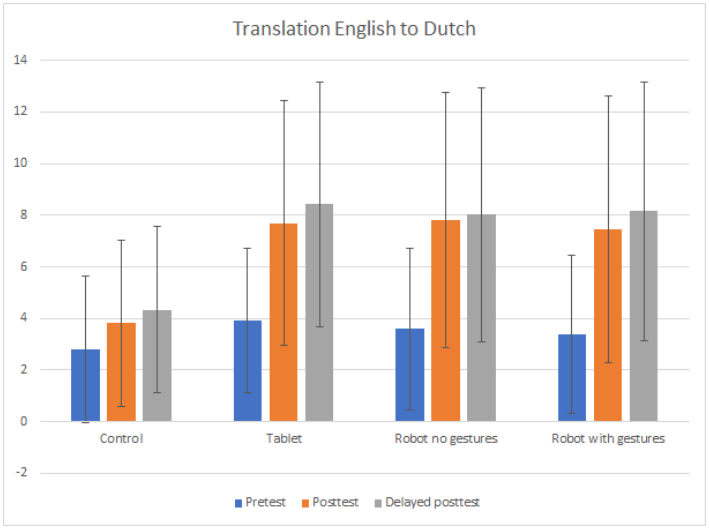
In this section we report the results of analyses conducted to test the three pre-registered hypotheses (see section 2 and deliverable 7.1, where these hypotheses are described in detail). A previous version of these results is included in a paper submitted to HRI (Vogt et al., 2018). The submitted version of this paper is attached in Appendix I. Please note that this is a temporary version, as, if accepted, the paper will be revised based on reviewer comments.

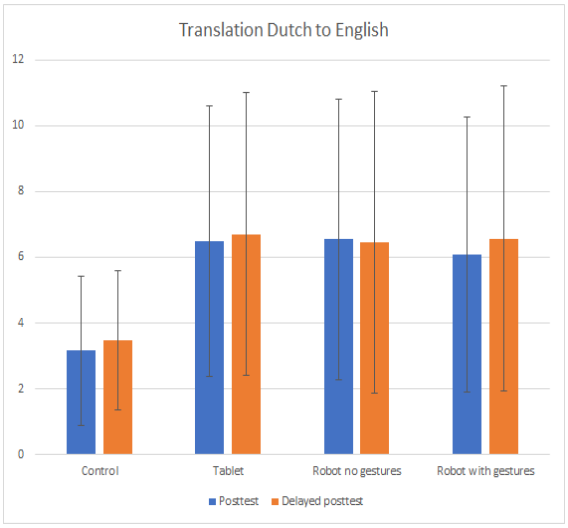
Figure 1 shows the scores for the three tasks across the pre-test and two post-tests. One sample t-tests were used to test whether children's sores on the pretest (translation task from English to Dutch; M = 3.5 words, SD = 2.96) were significantly higher than zero ($t(193) = 16.45$, $p < .001$, $d = 2.37$). Thus, during the pre-test children knew some of our target words. Note, however, that children on average knew only very few words and children who knew more than half of the words were excluded from participation ($n = 3$). See section 4 for a detailed description of which words children knew beforehand. Also when tested separately for the different conditions the results of the pretest were significantly higher than zero for all the different conditions (all p values < .001, d range = 2.02-2.85).

Scores of the translation tasks during the post-tests are also significantly higher than zero for all conditions (all p values<.001, d range = 2.42-3.60). The total scores on the translation tasks were, however, relatively low (range of mean scores across experimental conditions = 6.08-8.42 words translated correctly) compared to the maximum possible score of 34. A series of t-tests for paired samples showed that scores on the translation from English to Dutch test during the first post-test were significantly higher than scores during the pre-test for all conditions (for the experimental groups p values <.001, d range = 1.18-2.69, for the control group p = .008 and p = .012, d =1.02 and d = .94). Scores on the comprehension task were relatively higher (range 29.30-30.45 words correctly identified out of possible 54) and differed, in all conditions, significantly from the chance level score of 18 during the first post-test (all p values<.001, d for experimental conditions = 3.61-3.68, d control condition = 2.15) and the delayed post-test (all p values < .001, d for experimental conditions = 3.57-4.04, d for control condition = 2.69). Thus, in all tasks children show some knowledge of the target words and scores on the comprehension task are relatively higher than on the translation tasks. Notably, children in the control group also demonstrated knowledge of the target words during the post-tests. The effect sizes of the control group are, however, smaller than those of the experimental groups.

Several explanations are plausible for this growth. First, this can be the result of maturation. In the two months between the pretest and the first post-test children might learn English from other sources than our lessons, such as television and ambient speech. Therefore, the comparison of the control and experimental groups is important to indicate that children did learn during the lessons. Second, this might also be the result of spill over effects where children learned the words from peers who were in the experimental conditions. Finally, it is possible that children also learned from the test itself.

No clear differences were seen between the experimental groups. Given the high correlations between the translation tasks (see table 2, section 2) scores on these tasks were combined by computing the mean score. The comprehension task was maintained as a separate variable in the analyses as it correlated less strongly with the other tasks (though still very high), used a different format of testing, and tested a slightly different type of vocabulary knowledge.





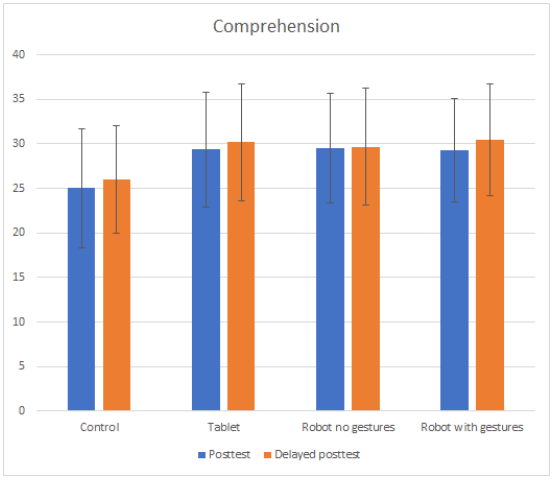


Figure 1. Scores on tasks measuring knowledge of target words across pre- and post-tests

To test our hypotheses, we performed a repeated measures analysis using a doubly multivariate design. This design entails a two (two post-test moments) by four (four conditions) design applied simultaneously to both outcome measures (the mean of the translation tasks and the comprehension task). Results showed a main effect of condition ($F(6, 378) = 3.34$, $p = .003$, $p\eta^2 = .05$). Post-hoc tests using the Bonferroni correction showed that scores in the experimental conditions were higher than in the control group (see Table 3 for results of these comparisons). No significant differences were found between the experimental conditions. Additionally, a significant main effect of time was also found ($F(2, 190) = 5.99$, $p = .003$, $p\eta^2 = .06$) showing that scores of the delayed post-test were higher than scores of the first post-test for both the translation tasks and the comprehension task.

Table 3
Univariate and post-hoc tests results of repeated measures analysis

	Translation tasks		Comprehension task	
	Univariate tests			
	F(df)	$p\eta^2$	*F*(df)	$p\eta^2$
Condition	6.58(3,190)***	.09	9.00(3,190)**	.07
Time	11.65(1,190)**	.05	4.01(2,290)*	.02
	Post-hoc pairwise comparisons			
	Mean difference (SE)	*d*	Mean difference (SE)	*d*
control - tablet only	3.63(.92)**	.89	4.29(1.28)**	.76
control- robot without gestures	3.51(.92)**	.86	4.08(1.28)**	.78
control - robot with gestures	3.41(.92)**	.83	4.45(1.28)**	.78
tablet only- robot without gestures	.12(.79)	.02	.21(1.11)	.03
tablet only- robot with gestures	.23(.79)	.06	.16(1.11)	.03
robot without gestures - robot with gestures	.11(.79)	.03	.37(1.11)	<.01

Note. * $p < .05$ ** $p < .01$ ** $p < .001$

Finally, a model was tested 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 task and the comprehension task during the first post-test as dependent variables. Results revealed that, besides the effect of condition already shown in the previous analyses, children with larger receptive vocabularies learned more English words. No significant effects of phonological memory and no interaction effects were found (see Table 4 for the exact results of these analyses). Analyses with the scores of the delayed post-test show the same trend for the translation tasks, although there was no effect of vocabulary on scores of the comprehension task during the delayed post-test.

Table 4
Results of regression analysis controlling for effect of Dutch vocabulary and phonological memory

	Translation tasks			Comprehension task		
	B(SE)	β	R^2	B(SE)	β	R^2
	First post-test					
Dutch vocabulary	.06(.02)	.16*	.03^{a}	.09(.04)	.17*	.02^{a}
Phonological memory	.08(.11)	.05		-.16(.16)	-.07	
Tablet only condition	3.80(.92)	.40***	.13^{b}	4.82(1.39	.34***	.09^{b}
Robot without gestures condition	3.62(.92)	.38***		4.67(1.39	.33***	
Robot with gestures condition	3.35(.91)	.35***		4.39(1.38)	.31**	
	Delayed post-test					
Dutch vocabulary	.05(.03)	.13†	.03^{a}	.05(.04)	.10	.01^{a}
Phonological memory	.11(.11)	.07		-.03(.17)	-.01	
Tablet only condition	3.70(.95)	.38***	.11^{b}	4.38(1.45)	.30**	.07^{b}
Robot without gestures condition	3.21(.95)	.33***		3.73(1.45)	.26**	
Robot with gestures condition	3.44(.94)	.35***		4.54(1.44)	.31**	

Note. $^{\dagger}p < .10$ $^{*}p < .05$ $^{**}p < .01$ $^{**}p < .001$ ^{a} value for a model with only vocabulary and phonological memory ^{b} value for complete model.

To summarise the findings described in this section, we found support for our first hypothesis: Results show that children can learn L2 words with a social robot. After the training, children in the experimental conditions with the robot knew more of the L2 target words than children in the control condition. We found no evidence in support of our second and third hypotheses. Children in the three different experimental conditions did not significantly vary in their knowledge of target words after completing the training, suggesting that learning with a robot and a tablet is not more effective than learning with a tablet only, and that the use of iconic gestures by the robot did not have an added value in our lessons. In section 7 (discussion) we further discuss the meaning of these findings and relate them to the findings described in other sections.

4 In-depth analyses of learning gains

To gain insight into which (kind of) words are learned better or worse by children throughout the lessons, in-depth descriptive analyses of children's learning gains were performed. More specifically, for each timepoint (i.e., pretest, post-test, delayed post-test) and each condition we identified for every word the percentage of children that knew that word, as measured by the translation tasks and the comprehension task. We decided to conduct this type of analyses in addition to the analyses of general tendency measures described in section 3, as it might be that certain conditions are beneficial for certain types of words but not for others. For example, for learning verbs, the use of gestures and active re-enactment might be beneficial (see for example Glenberg, 2008), while this might be less important for prepositions or number words. Moreover, these exploratory analyses will provide insights regarding which words were best learned in each condition. As the strategies used to teach words in the training vary somewhat between different words, these insights will enable us to see what strategies best worked and for what words in each condition.

Learning gains in individual words (or word categories) were measured as the change over time in percentage of children who demonstrate productive and/or receptive knowledge of a certain word. Children often perform better on word-knowledge tasks that are administered some time (e.g., a week or several weeks) after a vocabulary training than on tasks that are administered directly after the training. New words need time to become consolidated: they need to be integrated into children's memory and knowledge of these words needs to be strengthened (for a review, see Axelsson, Williams, & Horst, 2016). Sleep helps to strengthen and generalize this information, and thus plays an important role in word consolidation (Diekelmann, Wilhelm, & Born., 2009; Stickgold & Walker, 2013).

Children's learning gains were analysed at several levels. First, learning gains at the word level were analysed, to compare specific target words. Second, to check for patterns in learning gains that go beyond separate words, we clustered the target words into the following semantic categories (see also Table 5): movement verbs (n = 9), measurement words (n = 6), prepositions (n = 8), count words (n = 5), (mathematical) operations (n = 2), and comparatives (n = 4). These division in semantic categories is used in curricula designed to teach vocabulary to young children (see for example: https://www.gov.uk/government/collections/national-curriculum). As described in deliverable 1.1, the target words included in the study were chosen based also on such curricula, in addition to other indicators such as corpora and age of acquisition lists. Analyses at the semantic word category level enable us to test if the differences in learning gains are related to the semantics of the different words. For example, are movement verbs generally learned better with gestures while prepositions are better learned without? Is one type of words better learned without the robot then with the robot? What kind of words are learned by the control group? Learning gains were further also analysed at the lesson level and at the domain level (i.e., number and space domain), enabling us to examine possible differences between lessons and domains. Table 5 displays the target words sorted by word category, lesson, and domain.

Table 5

Overview of the target words sorted by word category, lesson, and domain

Domain/Lesson	**Target word**	**Word category**
Number domain		
Lesson 1: Zoo (1)	One	Count words
	Two	Count words
	Three	Count words
	More	Quantity words: Comparatives
	Add	Quantity words: Operations
	Most	Quantity words: Comparatives
Lesson 2: Bakery	Four	Count words
	Five	Count words
	Fewer	Quantity words: Comparatives
	Take away	Quantity words: Operations
	Fewest	Quantity words: Comparative
Lesson 3: Zoo (2)	Big	Measurement words
	Small	Measurement words
	Heavy	Measurement words
	Light	Measurement words
	High	Measurement words
	Low	Measurement words
Space domain		
Lesson 4: Fruit shop	On	Prepositions
	Above	Prepositions
	Below	Prepositions
	Next to	Prepositions
	Falling	Movement verbs
Lesson 5: Forest	In front of	Prepositions
	Behind	Prepositions
	Walking	Movement verbs
	Running	Movement verbs
	Jumping	Movement verbs
	Flying	Movement verbs
Lesson 6: Playground	Left	Prepositions
	Right	Prepositions
	Catching	Movement verbs
	Throwing	Movement verbs
	Sliding	Movement verbs
	Climbing	Movement verbs

The analyses were performed separately for each task: translation task L2>L1, translation task L1>L2, and the comprehension task. These analyses were not guided by specific hypotheses. Rather, as

mentioned above, this was an exploratory investigation. These analyses could provide us with additional insights about what specific words or word types are best learned under what conditions and about other factors that might affect word learning (such as the method used in each lesson). See Appendix II for mean percentages and standard deviations for analyses conducted at the level of individual words, semantic category, lesson and domain, for all conditions on all tasks. Below a summary of these results is given. Because of the different nature of the translation and comprehension tasks (measuring, respectively, translation ability in both directions and receptive knowledge), and the associated different manner of analysis, the analyses and results from these tasks are described separately.

Translation tasks

In the translation tasks, children heard each of the target words in English (L2>L1) or Dutch (L2>L1), and were asked to translate the word (i.e., what does [jump] mean in Dutch/English?). For each word the percentage of children who had given the correct translation of the word was calculated. Descriptive statistics are added in Tables 9-12 in Appendix II for the L2>L1 task, and in Tables 13-16 in Appendix III for the L1>L2 task. Below we will first report on children's scores at the pretest, and subsequently on the differences in word knowledge between the different timepoints, focussing respectively on the L2>L1 and the L1>L2 translation task.

Pretest. The pretest included only the translation task L2>L1. Mean scores show that children knew few English words during the pretest (*M* = 3.44, *SD* = 2.92). Analysis of this task at the word-level showed that count words were known by the highest percentage of children (i.e., one, two etc.; average group percentages for this word category ranged between 44% to 63% across the different conditions). As these were the words taught during the first two lessons, this vocabulary was expected to be learned most easily. The word category in which children scored the second best was movement verbs, especially the movement verbs taught in lesson 5, such as running and jumping. However, these words were only known by a very small group of children (3%-7% across the different conditions).

In general, the words that children knew the least in the pretest were operations and prepositions. For example, no children knew the operation words 'add' and 'take away', the word 'fewer' and the prepositions 'in front of' and 'behind'. The word that most of the children knew was 'five' (with group averages up to 83%). This was consistent across conditions and might be due to the high similarity between the word 'five' in English and in Dutch (vijf) and/or due to the fact that a lot of the participating children were five years old (children often mentioned that this was their age).

Pretest - post-test 1. The difference in performance between the pretest and the first post test is informative with respect to the learning gains. As mentioned in section 3, children score on average higher on the translation task L2>L1 conducted during the first post-test than during the pretest (*M* = 7.07, *SD* = 4.88). The results indicate that learning gains are largest for the word category movement verbs, followed by the count words. Learning gains per word were measured as the difference between percentage of children who knew the words during the pretest and the first post-test. The average learning gains for the experimental conditions were about 25% for the movement verbs (range 24%-25%), and about 12% for the count words (range 9%-15%). Prepositions and operations were the word categories with the lowest learning gains, with average learning gains ranging from 0% to 4% across conditions. Measurement words and comparatives fall in between with average learning gains between 7% and 11%. Thus, it seems that the semantic word categories that were familiar to the largest percentage of children beforehand, are also the categories that were best learned during the lessons.

At the word level, it can be seen that multiple words showed very low learning gains. These include words such as 'fewer', 'heavy', 'light', 'next to', and 'in front of'. The word 'add' appears to be very

difficult to learn in our lessons, as almost no children, in any condition, provided a correct translation during the post-test. The target words 'running', 'jumping', and 'sliding' showed the largest learning gains, as measured by the difference between the pretest and first post-test.

The analysis at lesson level indicated that the lessons from which children learned the most words were lesson 5 and 6, as these contained mostly (though not exclusively) movement verbs. The target words from lesson 2 reflected the smallest learning gains. This might be due to the fact that the word 'five', included in this lesson, was already familiar to most children during the pretest. The other four words taught in this lesson concern the word 'four', which showed learning gains in line with other count words, and the words 'take away', 'fewer' and 'fewest', which showed relatively low learning gains.

Overall, across conditions and timepoints, children score higher on the number domain than on the space domain but show larger learning gains in the space domain. There were no noticeable differences in the learning gains between the three intervention conditions at any level of analysis (in line with the statistical analyses described in section 3).

It should be noted that the control group also showed an increase for word knowledge of count words (8%) and movement verbs (4.5%), although these learning gains are considerably smaller. This is in line with the findings we reported in section 3, where we also offer several explanations for this effect.

L1>L2 task. The translation task L1>L2 was not included in the pretest, therefore no difference scores are reported. The scores on this task at the first post-test are rather low (M = 5.85, SD = 4.08). The patterns of performance at the word category, lesson and word level are similar to the patterns seen in the results of the L2>L1 task, although overall the percentages of children who knew each word (category) are lower for the L1>L2 task. To illustrate, about 20% of the children in the experimental conditions could translate movement verbs from Dutch to English, whereas about 30% could translate them from English to Dutch. It seems logical that retrieving a word in a foreign language as is done here is more difficult than identifying this word but retrieving a word in one's first language. Remarkably, the opposite holds with respect to the count words where it seems that children perform better on the L1>L2 task. When examined separately we see that for most count words this difference is rather small (2.5% to almost 10%), except for the word 'five' where about 18% more children get the translation from Dutch to English correct. This might be a bias in the way children responded in this task. When asked to provide English translations, children who did not know the answer exactly often repeated the target word they heard and only slightly adjusted their pronunciation. They did not do this when asked to provide Dutch translations. The count words, and especially the word 'five', are rather similar in Dutch and English. The experimenters who conducted the testing in our study were all Dutch speakers but not native English speakers. Therefore, it might be that when children simply repeated the target words they heard, it was easier for the experimenters to identify wrong pronunciation when children were just repeating the target words in the L2>L1 task then in the L1>L2 task.

Post-test 1 - post-test 2. The difference in performance between the first post-test and the delayed post-test is of relevance with respect to retention of the learned target words after a few weeks. Growth in performance between these two timepoints can be expected based on the so called 'consolidation effect' that is often found in word-learning interventions. This effect might be due to the positive effects of sleep on children's memory for recently encountered novel words (Axelsson, Williams, & Horst, 2016). The results indicated on average a small increase of performance for both the L2>L1 task (M_1 = 7.07, SD_1 = 4.88; M_2 = 7.57, SD_2 = 4.83) as the L1>L2 task (M_1 = 5.85, SD_1 = 4.08; M_2 = 6.04, SD_2 = 4.32). As described in section 3, these differences are statistically significant and apply to all conditions. Thus, it seems that in general a consolidation effect is present. However, it should be noted that the first post-test took place relatively long after the first lessons (i.e., at least 3-4 weeks). Therefore, the scores on the first post-test might also partly be caused by the consolidation effect. Moreover, it should be noted that the increase in scores between the first

and second post-test might also partially be explained by children learning from the test self. When looking at the word category, lesson, and word level, no clear patterns in growth or decline of performance are visible, as performance scores for some words or word categories increased for some conditions whereas it decreased for other words or in other conditions.

We can only speculate about why certain words or word categories are retained better in memory than others. The literature also does not give clear explanations for this. It might be that for some words were linked to more elaborate or rich semantic or conceptual maps. For example, the active performance of movement verbs might have enabled a more elaborate concept for these words (see also Glenberg, 2008), that, in turn, supports retention of these words. In addition, it is possible that words that are more frequent in ambient speech, such as the count words, are better retained as children were in a way primed for these words. The decrease in word knowledge of certain words might be due to the complexity of the concepts, such as mathematical operations. It is possible that these concepts are not completely anchored in children's native language, which makes it extra difficult to remember these words in another language. It is very likely that multiple factors are simultaneously at play. These possibilities point to interesting venues for future analyses and research.

Comprehension task

In the comprehension test, children saw three photos or films simultaneously on a laptop screen and were asked to choose which image matches a target word. Only data from children who had given an answer to at least one third of the trials was included. Data from trials where children answered 'I do not know' was considered missing. While in the translation task such an answer was a viable option, in the comprehension task children can guess the answer (as they usually did when they did not know the answer), therefore this answer could likely imply that children did not understand the task or point to non-compliance. Data of three children were excluded from the first post-test, and data of three different children were excluded from the delayed post-test as they failed to answer at least one third of the trials. As described in the introduction, a selection of 18 target words was included in the comprehension task, with each word being tested using three different trials (i.e., in total 54 items). For the current analyses, for every word the number of trials where children provided a correct answer was calculated per word (i.e., word score), which could thus range from 0 to 3. If a child had given no answer to one of the three trials, no word score was calculated for that word for that child. A score of 2 or higher was seen as an indication of word knowledge. A score of 1 was considered as guessing. Descriptive analyses are presented in Tables 17-20 in Appendix IV. Since the comprehension task was also not included in the pretest, we focus on the performance scores at the first post-test and at the difference in word knowledge between the two post-tests.

Post-test 1. The results of the first post-test display in general the same pattern as described above for the translation tasks (M = 28.80, SD = 6.38). The average percentages of children who know the words (i.e., answer correctly on at least 2 trials) in the different experimental conditions range from 70% to 76% for the movement verbs and from 73% to 85% for count words, which were again the word categories that the children knew best. Prepositions and operations were again the word categories that children scored the lowest on, yet average group percentages were about 27% for operations and ranged from 32% to 40% for prepositions, which is considerably higher than in the translation tasks (these are the percentage children of word score 2 and 3 together). This might be because receptive knowledge is often more easily acquired than productive knowledge (Mondria & Wiersma, 2004). Moreover, the images used in this task might remind the children of what the words mean, whereas in the translation tasks children did not receive any clues for the meaning of the words.

Post-test 1 - post-test 2. The retention of learned target words in the comprehension task is also comparable to the translation tasks. The mean performance score on the second post-test was slightly higher than on the first post-test (*M* = 29.65, *SD* = 6.38), which might be because of the consolidation effect. Again, no clear patterns could be derived from the change in performance between the two post-tests, with differences in performance scores ranging from -7% to +11% between the two post-tests at the semantic word category level. The differences between the two post-tests were in general somewhat bigger in this task compared to the translation tasks.

Taken together, the in-depth analyses of the translation tasks and comprehension task display in general the same pattern of results. Children's learning gains are especially noticeable for count words and movement verbs, and the learned word are generally retained. The results indicate that it is important to consider word effects, as there are differences in the learning gains at the word, category, lesson, and domain level.

5 Immediate learning gains

In addition to the post-tests conducted after the lesson series, children's knowledge of the target words was also tested at the end of each lesson. After all the target words of that lesson were taught, children had to complete a short task to examine their knowledge of the novel target words learned during that lesson. Children were presented with 3 images on the tablet screen and were asked to click on the image that corresponded to the target word they heard. Each target word was tested using two different trials with different sets of images. The animations used for the items were similar to the animations used during the lessons. Children did not receive any feedback in this task and the task proceeded once they chose an image (whether it was correct or not). We refer to children's performance at this task as 'immediate learning gains'.

It is interesting to study changes in these immediate learning gains as they may provide insights regarding the, so called, novelty effect. This effect denotes an initial improvement in performance because of high interest in a new stimulus, in this case the social robot, and decline of performance over time due to lower interest after becoming familiar with the social robot (Kanda, Hirano, Eaton, & Ishiguro, 2004). This suggests that learning gains might decrease as novelty wears off (Leite, Martinho, & Paiva, 2013). Immediate learning gains were analysed at the lesson level and at the word level, and the results of these analyses are described below. Moreover, a comparison of these results with the results provided in the previous sections can show whether the words best retained over time are also learned well initially. Note that in this analysis only data from the experimental conditions were included, as the control group did not follow the lessons and therefore, did not do these tests.

Proportion of the trials answered correctly were calculated per lesson. Results are presented in Table 6, and Figure 2 provides a visualisation of the changes in the immediate learning gains over time. Task performance in all conditions decreases steadily over time, with a (relatively) sharp increase in the last lesson. A One-Way Repeated Measures ANOVA was performed to test the change in performance scores over time. Time was included as within-subjects factor and condition as between-subjects factor, to examine whether conditions differ in change in performance. The results indicated a main effect of time, F (5, 795) = 3.678, $p < .001$, partial $\eta^2 = 0.15$. Post-hoc tests were performed using the Bonferroni correction. The results are displayed in Table 7. No significant differences were found between the conditions, F (1, 159) = 0.373, p = 0.690. There was also no interaction effect between time and condition, F (10, 795) = 0.319, p = .265.

At the word level, the number of correctly answered trials was calculated per child, which could range from 0 to 2. Only scores of 2 were seen as an indication of word knowledge. Detailed results are shown in Table 21 in Appendix V. The general pattern of learning gains is comparable with those from the translation and comprehension tasks used at the post-tests, as described in section 4, with large differences in learning gains between words.

Table 6
Means and standard deviations of proportion correctly answered trials per lesson for each condition separately and for the total intervention group

Lesson	Tablet-only (n = 54)		Robot with iconic gestures (n = 54)		Robot without iconic gestures (n = 54)		Total intervention group (n = 162)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	0.56	0.21	0.61	0.21	0.58	0.18	0.59	0.20
2	0.54	0.20	0.54	0.18	0.54	0.19	0.54	0.19
3	0.51	0.21	0.49	0.20	0.51	0.21	0.51	0.21
4	0.47	0.21	0.42	0.19	0.45	0.18	0.45	0.19
5	0.43	0.24	0.37	0.19	0.39	0.21	0.40	0.21
6	0.55	0.17	0.52	0.13	0.48	0.18	0.52	0.17

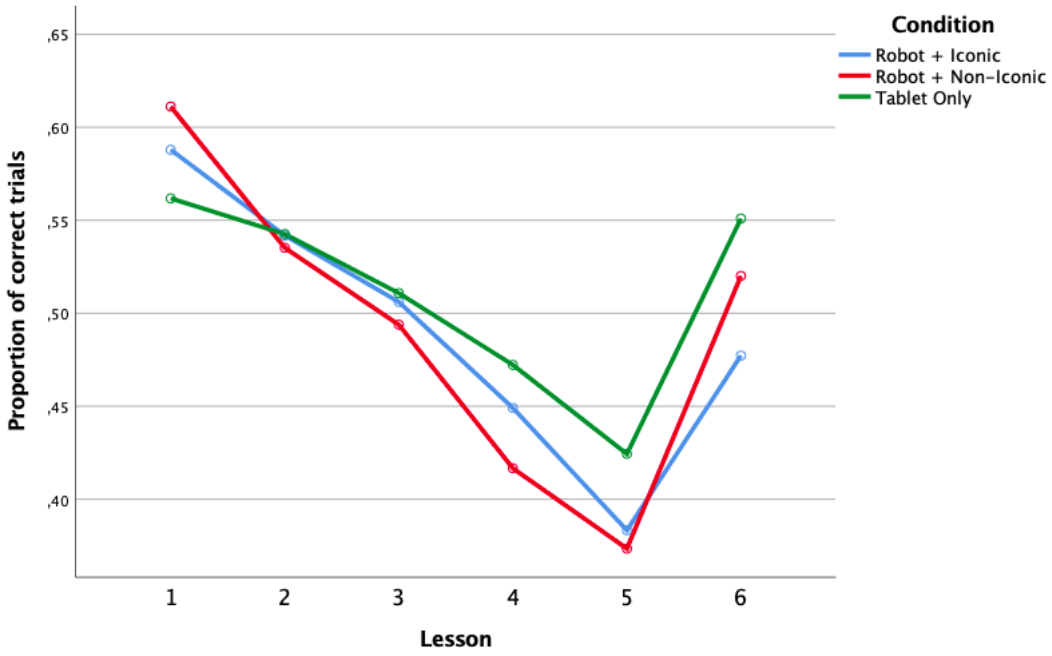


Figure 2. Changes in proportion of correct trials at the end of each lesson

Table 7

Pairwise comparisons of scores at the end of the lessons

Lesson (A)	Lesson (B)	Mean Difference (A-B)	SE	*p*	*d*
1	2	0.047	.015	.040	0.24
	3	0.080	.017	.000	0.39
	4	0.138	.018	.000	0.70
	5	0.190	.017	.000	0.91
	6	0.070	.017	.001	0.38
2	3	0.034	.018	.883	0.17
	4	0.091	.018	.000	0.48
	5	0.143	.018	.000	0.71
	6	0.023	.017	1.000	0.13
3	4	0.058	.018	.030	0.29
	5	0.110	.018	.000	0.52
	6	-0.010	.020	1.000	-0.05
4	5	0.052	.018	.063	0.26
	6	-0.068	.019	.009	-0.38
5	6	-0.120	.018	.000	-0.63

In lesson 1, children scored relatively high on the count words, and subsequently on the comparatives ‘more’ and ‘most’. It turns out that add was very difficult for the children, as only 7% of the children knew this word at the lesson test. The relatively high lesson score of lesson 2 is mostly due to high scores on the count words. Similar to the results of the pre- and post-tests, 96% of the children knew the word ‘five’. In lesson 3, the word ‘small’ was known by relatively many children (69%), followed by ‘big’ and ‘light’. In contrast, the measurement word low appears to be quite difficult (known by 12% of the children).

In lesson 4, scores on the word ‘falling’ are considerably higher than the other words, which were all prepositions (52% whereas the other words score on average 20%). The preposition ‘on’ appears to be very difficult and was answered correctly only by 6% of the children. This is surprising, as this is one of the

simpler, first learned prepositions, at least in L1 (Brysbaert, Stevens, De Deyne, Voorspoels, & Storms, 2014). One possible explanation is that in lesson 4 children were asked for the first time in the training to listen to and repeat an entire phrase in L2 (e.g., 'the apple is on the table'). 'On' was the first target word introduced in lesson 4. It might be that children still needed to get used to this new way of introducing and learning words and therefore did not learn the word 'on' that well. Alternatively, it might be that children were confused by the word 'on' since in Dutch there are two different prepositions representing slightly different meaning of the English preposition 'on'. In Dutch 'op' refers to being on top of something in a horizontal position (e.g., op de tafel [on the table]) whereas 'aan' refers to 'on' in a vertical position (e.g., aan de muur [on the wall]).

The relatively low lesson score of lesson 5 is mostly due to the words 'in front of', 'walking', and especially 'behind', which only few children knew (5%). The best known word of this lesson was 'flying', which almost half of the children knew (47%). In the final lesson, lesson 6, children scored high on 'sliding' (70%) and climbing (50%). The words 'catching' and 'throwing', and 'left' and 'right' were considerably more difficult (16%-21%).

Finally, comparing scores at the test conducted at the end of each lesson with the scores at the first post-test allowed us to examine differences between immediate recall and retention at the end of the lesson series. These scores might differ since new words need time to become consolidated (Axelsson, Williams, & Horst, 2016). It should be noted that unlike retention measured as the difference in scores between the first and the second post-test, the time gap between learning the words in the lessons and the first post-test varies between words and between children. The order of the lessons was always identical, but the length of the entire lesson series and the number of lessons per week varied within pre-set ranges (see deliverable 7.1). As the pattern of results for the second post-test was highly similar to that of the first, no additional comparisons with the results of this second post-test were conducted.

We compared the scores of the lesson tasks with the comprehension task, as these tasks are similar in nature. The comparison was only possible for the words that were included in the comprehension task. The rank order of scores seen in the lesson tasks was comparable to the order seen in the comprehension task, regardless of condition. However, children scored higher on the comprehension task during the first post-test than on the lesson task. The difference in scores varied between words. For some words the difference was small, but for words such as, 'heavy', 'add', and 'catching' the difference was about 30% to 40% of children who knew the words during the first post-test but not during the lesson task. For 'jumping' and 'on', the differences even go up to 55%. These results provide an additional indication for the idea that it takes time for new words to become consolidated, in line with Axelsson, et al., (2016) as described above.

In future analyses of these data we will check, using mixed linear models, whether different factors influence performance at the task at the end of the lessons, such as duration of the lessons, time between the lesson, individual characteristics of the children and words learned.

6 Perception of the robot and learning gains

In this section we report the results of an analysis conducted with data regarding children's perception of the robot. These results were included in a paper submitted to HRI (van den Berghe, de Haas, et al., submitted; the full paper is attached in Appendix VI). Please note that this is a temporary version. If accepted the paper will be revised, based on reviewer comments.

When people interact with a social robot they tend to attribute human form, characteristics and/or behaviours to the robot. This phenomenon is called anthropomorphism (Bartneck, Kulic, Croft, & Zoghebi, 2009) and is a well known phenomenon regarding many other types of objects. Anthropomorphism might be useful as it might enable children view the robot positively and collaborate with it better than if they do not view it as a human. See Appendix VI for a detailed review of individual differences in the development of this phenomenon.

In the large-scale evaluation study, we measured children's anthropomorphism of the robot during the pre-test and the first post-test. Note that prior to the pre-test children had seen the robot once during the group introduction session. Anthropomorphism was measured using a questionnaire administered by an experimenter that took about 5 to 10 minutes to complete. The questionnaire was based on the work of Jipson and Gelman (2007). The questionnaire contained 19 questions divided in seven questions about the 'biological' properties of the robot (e.g., can it break?), seven questions about the mental state properties of the robot (e.g., can it be happy?), one question about the gender of the robot, and another four questions determining the role of the robot as a peer or a robot and one general question about whether children see the robot as a human or a machine. See table 8 for a list of all questions. Each question could be answered with yes/no/I do not know. When children answered yes or no they were asked to explain their answers. In this report we only included data about children's yes/no answers. Data about their explanations will be discussed in following deliverables.

Table 8. Items of the perception questionnaire

Biological properties	Mental state properties	Other aspects
Do you think Robin the robot...		
can see things?	can be sad?	is a boy or a girl?
is made by a human?	can remember something?	is more a teacher or a friend?
can feel it when you tickle?	can think?	is more a friend or a thing?
has to eat?	understands when you say something?	is more a teacher or a thing?
can feel pain?	can enjoy something?	is more a human or a thing?
grows?	can be happy?	
can break?	can recognise you?	

One point was given to each answer that indicated anthropomorphism of the robot. For example, when asked if the robot has to eat, answering yes was awarded one point and answering no was awarded no points. The total score was the proportion of questions answered with human-like perception. Separate scores were computed for the biological and mental states items.

Detailed results can be found in Appendix VI. Here we provide a short summary. Results show that on average children perceived the robot more as a human than as a machine. Children ascribe more mental states properties than biological traits to the robot. When scores on the perception questionnaire were compared between the pre-test and post-test we saw that most children were stable in their anthropomorphism. However, considerable groups of children showed an increase or decrease in level of anthropomorphism. Further analysis showed that, as a group, children mostly changed their perception of biological properties. Children for example stopped believing that the robot could feel pain. Children in the two robot conditions did not significantly differ in their perception of the robot nor in the change in perception over time.

Finally, the relation between anthropomorphism and learning gains was studied. Results showed that children who anthropomorphized the robot more during the pretest learned less than children who saw it more as machine. Moreover, an increase in the degree to which children anthropomorphized the robot was related to higher performance on the comprehension task during the delayed post-test. See Appendix VI for an extensive discussion of these results.

In addition to the results reported in van den Berghe, de Haas, et al. (submitted) we also tested whether children perceived the robot more as male or female. In both the pre- and the post-test, the majority of children (89%) perceived the robot as male. During the pre-test the perception of the robot's gender does not significantly vary between boys and girls. In fact, out of 20 children who perceived it as a girl, exactly 50% are girls. During the post-test, however, 17 out 23 children who perceived the robot as a girl were girls themselves. Chi square test that this difference is significant ($\chi^2(1)=5.46$, $p=.019$, $\varphi=.17$).

During the pre-test, most children saw the robot as a friend rather than a teacher (86%) and this perception is also seen in the post-test where 78% of the children see the robot as a friend. This suggests that our framing of the robot as a peer tutor was successful. Interestingly, when directly asked if the robot is a person (either friend or teacher) or a thing, during the pretest children's answers are divided between person and thing (percentages of children perceiving the robot as a thing across the questions range from 48% to 59%). When asked these questions again during the post-test, most children indicate to perceive the robot as a thing (percentages range across the questions 51%-74%).

In future analyses we will examine the explanations children provide for their answers on the questions to further gain insight in children's perception of the robot and the way it might facilitate or impede word learning. Nevertheless, the current results suggest that the perception of the robot might play a significant role in the learning process and is definitely something to be taken into account when designing robot tutors.

7 Discussion and conclusion

This deliverable reports on first analyses we conducted on the data of the large scale L2TOR evaluation study. As far as we know, this is the first study to test the effect of an (almost entirely) autonomous social robot on L2 word learning over a longer period of time. In this section we will discuss a few of the main findings and sketch our plans for future analyses of the data. Moreover, unlike many studies in social-robotics in education, the current study compares learning with the robot to children's learning without any intervention (control condition) and to learning using touch screen-based technology. These comparisons are crucial for the implementation of robots in education. It is clear that children learn L2 (English in our case) also without explicitly being taught. Therefore, every new training programme should always be compared to a control group. Moreover, given the costs of a novel technology such as social-robots, educational institutions will only consider implementing such technologies if they have a clear added value beyond that of already used, cheaper, technology.

Results show that children can learn L2 words using the training programme we designed and that they retain these words over a period of 3 to 5 weeks after the training ends. However, the presence of the robot and the use of iconic gestures do not make a difference for their learning. Though children learn new words, the effects are relatively small especially compared to other vocabulary training programmes (Marulis & Neuman, 2010). Due to the current limitations of robot technology, the training had a very static nature, meaning that children followed more or less identical lessons. That is, lessons could not be adapted to the level of the specific child. This static structure was chosen for several reasons. First, it turned out that the ASR for children is not reliable enough to use in the training (Kennedy et al., 2017) . Therefore, in order to enable the robot to function autonomically without a speech recognition system the possibilities for adaptations on individual level were restricted. Children received varying amount of feedback depending on their performance, but the content of the lesson and method of teaching were the same for all children. Second, given the systematically controlled experimental design of this study, we could only truly compare the different conditions if children received the same lessons. Most vocabulary training programmes involve a human teacher who usually adapts to the needs and pace of each child and can adjust the teaching strategy and way the information is presented according to the needs of specific children. Moreover, a human tutor responds contingently to the behaviour of children. To be truly effective, a robot tutor will need to be able to do this too. Within the current state of technology, it is still challenging for autonomous robots to achieve personalised adaptation. However, some work done within L2TOR (WP 5) provides demonstrations of how this adaptation might be achieved (De Wit et al., 2018; Schodde, Bergmann, & Kopp, 2017).

Given the current state of technology we could not avoid the use of a tablet. But it is very likely that the tablet impedes a more natural interaction between the child and the robot, that might do more justice to the potential of the robot as an embodied agent. One of the main limitations we had to deal with was the lack of automatic speech recognition (Kennedy et al., 2017). To solve this, as well as the difficulties related to object tracking (Wallbridge et al., 2017), while still enabling the system to function autonomously, we resorted to working with a tablet that mediated the interaction and provided the educational context. It is important to note that in our study the content was displayed on the tablet and the tablet was also used to record the child's responses. In the two robot conditions, the robot provides verbal support (i.e., instructions, translations, feedback) and non-verbal support (i.e., deictic gestures and, in one condition, iconic gestures). In the tablet only condition, the robot was hidden from sight, but the verbal support provided was identical to that in the robot conditions (the robot's voice was directed through tablet's speakers). The only difference, therefore, is that in the tablet only condition children did not receive the non-verbal support and there was no

robot present. Also in the robot conditions children could follow the training successfully also if their interaction with the robot was minimal, as they usually had to provide answers using the tablet.

While, based on previous work (e.g., de Wit et al. 2018) we assumed that non-verbal support would increase learning gains, it might be that the tablet was so engaging that it dominated the interaction. In the field it was clear that most children were primarily focussed on interacting with the tablet and playing the game that provided the educational context. One possible hypothesis is that, in the set up used in the current study, the robot might have even distracted the children from interacting with the task on the tablet. We are currently working on coding children's engagement with the task and with the robot to see if these patterns of engagement vary between the conditions and if they are related to learning outcomes. It might be that children in the tablet condition are more engaged all together, or that they benefit from the fact that they do not need to shift their attention between the robot and the tablet.

In the future, when ASR and object recognition technologies will be further advanced, it might be possible to design a programme where the interactions are more natural and embodied and make use of the real 3D world, rather than 2D screens. We previously showed that, in a set up like the one used in the current study, the use of physical objects did not show any advantage above manipulating objects on the tablets, when human tutors follow a strictly controlled script (Vlaar et al., 2017). However, it is possible that in more natural and adaptive interaction, the use of real-life objects might enable a deeper level semantic processing and will thus facilitate higher learning gains (Antonucci & Alt, 2011; Ernst, Lange, & Newell, 2007). However, within the current state of technology, this remains a hypothesis.

The specific use of iconic gestures did not contribute to children's learning. In Vogt et al. (submitted; see Appendix I) we discuss possible reasons for this lack of effect. In short, this might be due to the physical limitations of the robot that made the gestures sometimes look odd. Another possible reason is that we introduced the gestures too often (with every repetition of the target word) which might have ended up distracting the children. We are currently investigating the data at the word-level to see if some gestures worked better than others. We are also specifically looking if active re-enactment by the children, which we included for words like running and jumping, might have had a positive effect on learning gains (see also Glenberg, 2008).

Effects found in the field of social-robotics are often attributed to a so-called novelty effect where children learn from the robot because the novelty of interaction with it captures their attention and motivates them (Kanda et al., 2004; Leite et al., 2013). As in this study children interacted with the robot over a longer period of time we hoped to gain insight into a possible decrease in learning gains over time, as this effect would suggest. The findings presented in section 5 do not yet provide a clear answer to this question. When means are examined, it would seem that performance does decrease to almost chance level at lesson 5, in line with a hypothesis based on the novelty effect idea. But then in lesson 6 we do see an increase in learning gains. Moreover, comparison of learning gains across lessons is complicated as in every lesson different words are taught. The results at the level of single words show that also at the end of the lesson some words are learned better than others. Also in the first lessons, where on average learning gains as measured by the task at the end of the lesson are high, some words show very low performance. Similarly, in the later lessons where immediate learning gains are, on average, low, performance on some of the words is good. Moreover, when studying the results of the post-tests in terms of lessons (section 4) we do not see that words learned in later lessons are retained less well than words learned in earlier lessons. Thus, we do not find evidence for a clear novelty effect, but the design of the study involving different words per lesson, does not enable us to draw strong conclusions regarding this effect. However, we do plan to analyse these data further using mixed effects modelling where the words can be included as random factors. This analysis will enable us to account for the fact that learning gains may differ between words,

This difference in performance between the immediate recall and retention during the first and second post-test could also be due to differences in the way children learn words. While the immediate post-test conducted at the end of each lesson (i.e., the 'lesson task') involves short-term memory, retention in the first and second post-tests involves long term memory. A large body of psychological literature discusses the differences between these two types of memory (e.g., Axelsson et al., 2016; Cowan, 2008). Immediate recall might be influenced by factors such as recency (words last heard are remembered better), the distractors used for measuring the knowledge of specific words (e.g., children might be inclined to perform better when images are moving, or they might like certain images like a slide more than others) and so forth. Retention, on the other hand, usually reflects knowledge that has been consolidated and is less affected by temporary factors. The results of the analyses at the word level presented in sections 4 and 5 show that for some words the performance in immediate recall tests and in the long-term retention tests are comparable, but for some words we do see a change. Some words show the decay, as one can expect, while others seem to be remembered better during the post-tests than during immediate recall tests (right at the end of the lesson). It might be that some factors, such as for example active re-enactment, enable better retention over time as children create a richer semantic map for these words. This effect is not always seen during immediate recall which is influenced by other, more temporary, factors. In future analyses we will also compare learning gains at the word level to try to identify factors that impede or support long-term retention which is ultimately the goal of vocabulary training programs.

The use of iconic gestures did not seem to affect children's perception of the robot. However, as previously mentioned, the gestures did not look very natural, due to the physical limitations of the robot. Moreover, as we worked with a set script and standardised the number of gestures presented to be able to compare the children, the gestures were not used in a natural manner. Human speakers would vary more in gestures and would sometimes use and sometimes not use iconic gestures, depending on their own needs and the needs of their conversational partner. Thus, the use of iconic gestures in this case does not necessarily make the robot more human. However, as anthropomorphism seems to be related to learning it is worthwhile to study how to promote the perception of the robot as human. It is also important to study the individual differences in anthropomorphism to see what children are more or less likely to anthropomorphize and possibly adjust the training to this tendency.

In sum, in this deliverable we report the first results showing that while children learned L2 words in our training, there was no advantage of using a robot (with or without iconic gestures) over using just a tablet. We offer possible explanations for this finding and will conduct further in-depth analyses to explore learning gains in the different conditions and include possible moderators (such as vocabulary in L1, phonological memory, selective attention) to further explore individual differences. In order to be able to design a truly effective autonomous social-robot that can assist in L2 tutoring with this age group, some technological advances are required (i.e., ASR, adaptivity, object recognition). In Deliverable 7.3 we further discuss the implications of these findings for the field of education and translate these conclusions to practical recommendations.

Reference List

Antonucci, S. M., & Alt, M. (2011). A lifespan perspective on semantic processing of concrete concepts: Does a sensory/motor model have the potential to bridge the gap? *Cognitive, Affective, & Behavioral Neuroscience, 11,* 551-572. doi:10.3758/s13415-011-0053-y

Axelsson, E. L., Williams, S. E., Horst, J. S. (2016). The effect of sleep on children's word retention and generalization. *Frontiers in Psychology, 7*(1192). doi:10.3389/fpsyg.2016.01192

Bartneck, C., Kulic, D., Croft, E., & Zoghebi, S. (2009). Measurement instruments for the anthropomorphism animacy, likability, perceived intelligence and perceived safety of robots. *International Journal of Social Robotics, 1,* 71-81. doi:10.1007/s12369-008-0001-3

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica, 150,* 80-84. doi:10.1016/j.actpsy.2014.04.010

Chiat, S. (2015). Non-word repetition. In S. Armon-Lotem, J. de Jong & N. Meir (Eds.). *Methods for assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 227–250). Bristol: Multilingualism Matters.

De Wit, J., Schodde, T., Willemsen, B., Bergmann, K., De Haas, M., Kopp, S., Krahmer, E., & Vogt, P. (2018). “The effect of a robot’s gestures and adaptive tutoring on children’s acquisition of second language vocabularies” (pp. 50–58), in Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, ACM .

Diekelmann, S., Wilhelm, I., & Born, J. (2009). The whats and whens of sleep-dependent memory consolidation. *Sleep Medicine Reviews, 13,* 309–321. doi:10.1016/j.smrv.2008.08.002

Duffy, B. R. (2003) Anthropomorphism and the social robot. *Robotics and Autonomous Systems, 42,* 177-190. doi:10.1016/s0921-8890(02)00374-3

Ellis, N. C. (2006). Selective Attention and Transfer Phenomena in L2 Acquisition: Contingency, Cue Competition, Salience, Interference, Overshadowing, Blocking, and Perceptual Learning. *Applied Linguistics, 27*(2), 164–194. doi:10.1093/applin/aml015

Ernst, M. O., Lange, C., & Newell, F. N. (2007). Multisensory recognition of actively explored objects. *Canadian Journal of Experimental Psychology, 61,* 242-253. doi:10.1037/cjep2007025

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics, 19,* 537– 552. doi:10.1017/s0142716406060383

Glenberg, A. M. (2008). Embodiment for education. In P. Calvo, & T. Gomila. (Eds.), Handbook of cognitive science: An embodied approach (pp. 355–372). Amsterdam: Elsevier.

Jipson, J. L., & Gelman S. A. (2007). Robots and rodents: Children's inferences about living and nonliving kinds. *Child Development, 78,* 1675-1688. doi:10.1111/j.1467-8624.2007.01095.x

Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human-computer interaction, 19*(1), 61-84. doi:10.1207/s15327051hci1901&2_4

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., & Belpaeme, T. (2017). "Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations" (pp. 82-90) in Proceedings of the 12th ACM/IEEE Int Conf on Human-Robot Interaction, ACM.

Leite, I., Martinho, C., & Paiva, A. (2013). Social robots for long-term interaction: A survey. *International Journal of Social Robotics, 5*(2), 291–308. doi:10.1007/s12369-013-0178-y

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of educational research, 80*(3), 300-335. doi:10.3102/0034654310377087

Masoura, E. V., & Gathercole, S. E. (2005). Contrasting contributions of phonological short-term memory and long-term knowledge to vocabulary learning in a foreign language. *Memory, 13*(3–4), 422–9. doi:10.1080/09658210344000323

Mondria, J. A., & Wiersma, B. (2004). Receptive, productive, and receptive productive L2 vocabulary learning: What difference does it make. *Vocabulary in a second language: Selection, acquisition, and testing, 15*(1), 79–100.

Mulder, H., Hoofs, H., Verhagen, H., Van der Veen, I., & Leseman, P. P. M. (2014). Psychometric properties and convergent and predictive validity of an executive function test battery for two-year-olds. *Frontiers in Psychology, 5*, 733-2014. doi:10.3389/fpsyg.2014.00733

Schodde, T., Bergmann, K., & Kopp, S. (2017). "Adaptive robot language tutoring based on bayesian knowledge tracing and predictive decision-making," in Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction, ACM.
Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: Evolving generalization through selective processing. *Nature Neuroscience*, *16*(2), 139–145. doi:10.1038/nn.3303

Verhagen, J., & Leseman, P. P. M. (2016) How do verbal short-term memory and working memory relate to the acquisition of vocabulary and grammar? A comparison between first and second language learners. *Journal of Experimental Child Psychology, 141*, 65-82. doi:10.1016/j.jecp.2015.06.015.

Vlaar, R., Verhagen, J., Oudgenoeg-Paz, O., & Leseman, P. (2017). "Comparing L2 word learning through a tablet or real objects: What benefits learning most?", in Proceedings of the Workshop R4L, ACM/IEEE HRI.

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

Werker, J. F., & Yeung, H. H. (2005). Infant speech perception bootstraps word learning. *Trends in cognitive sciences, 9*, 519-527. doi:10.1016/j.tics.2005.09.003

Appendix I: Vogt et al., submitted

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

***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 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.

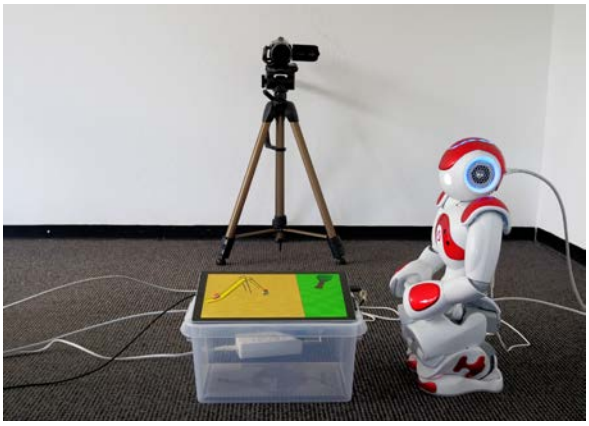


Fig. 1. The basic setup for all lessons.

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.

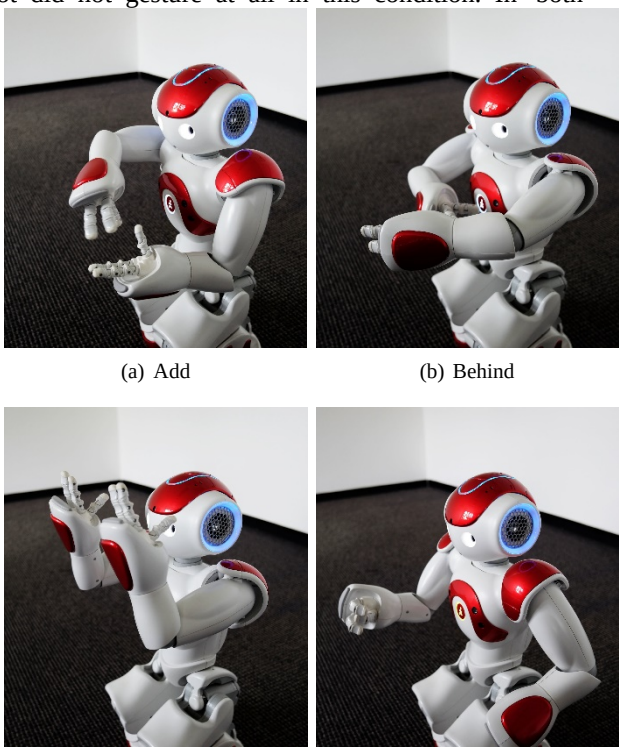
The seventh session was a recap lesson, where children created a picture book. They saw, one by one, the scenes of the six content lessons, and 'stickers' with the objects of these lessons. They placed these 'stickers' on the scenes, while the robot discussed with the children the target words that they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

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



(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.

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.

2) 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.

3) 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 N_b/N_g	Avg Age + SD (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) x 2 (post-tests) MANOVA with the three measures at the post-tests as dependent variables. The findings showed a main effect of condition ($F(9, 452.8) = 2.16, p = .023, \eta^2 =$ .034). Post-hoc tests (Bonferroni) showed that children in the exper- imental 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 ($F(3, 186) = 5.00, p = .002, \eta^2 =$ .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 systemat- ically 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 tothe 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 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 Interven- tion 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 lan- guage 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. Pa- padopoulos, 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 vocabu- lary test-III-NL*. Amsterdam: Pearson, 2005.

[29] H. Mulder, H. Hoofs, J. Verhagen, I. van der Veen, and P. P. M. Lese- man, "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 Tu- toring 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 *Proceed- ings 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.

Appendix II: Results of in-depth analyses: Translation task L2>L1

D = difference
PL1 = posttest 1
PL2 = posttest 2

Table 9
Translation task L2>L1: Control condition

Lesson	Target word	% correct pretest n=32	% correct pL1 n=31	% correct pL2 n=31	D pretest>pL1	D pL1>pL2
Number domain						
1	One	53.1	65.6	65.6	12.5	0
	Two	31.30	43.8	56.3	12.5	12.5
	Three	43.8	50	62.5	6.2	12.5
	More	6.3	9.4	9.4	3.1	0
	Add	0	0	0	0	0
	Most	0	3.1	9.4	3.1	6.3
2	Four	34.4	37.5	53.1	3.1	15.6
	Five	59.4	65.6	78.1	6.2	12.5
	Fewer	0	0	0	0	0
	Take away	0	3.1	0	3.1	-3.1
	Fewest	0	0	0	0	0
3	Big	3.1	0	0	-3.1	0
	Small	3.1	12.5	3.1	9.4	-9.4
	Heavy	0	0	0	0	0
	Light	3.1	6.3	6.3	3.2	0
	High	0	6.3	3.1	6.3	-3.2
	Low	0	0	0	0	0
Space domain						
4	On	3.1	0	0	-3.1	0
	Above	0	0	3.1	0	3.1
	Below	0	0	0	0	0
	Next to	0	0	3.1	0	3.1
	Falling	0	3.1	0	3.1	-3.1
5	In front of	0	0	0	0	0
	Behind	0	3.1	0	3.1	-3.1
	Walking	3.1	3.1	3.1	0	0
	Running	6.3	12.9	18.8	6.6	5.9
	Jumping	6.3	18.8	34.4	12.5	15.6

	Flying	6.3	15.6	12.5	9.3	-3.1
6	Left	6.3	0	0	-6.3	0
	Right	6.3	6.3	0	0	-6.3
	Catching	0	0	3.1	0	3.1
	Throwing	0	3.1	0	3.1	-3.1
	Sliding	3.1	9.4	3.1	6.3	-6.3
	Climbing	3.1	3.1	6.3	0	3.2

	Pretest		*PL1*		*PL2*		Mean D pre>pL1	Mean D pL1>pL 2
	M	SD	M	SD	M	SD		
Movement verbs (9)	3.13	2.73	7.68	6.70	9.03	11.33	4.54	1.36
Measurement words (6)	1.55	1.70	4.18	5.11	2.08	2.56	2.63	-2.10
Prepositions (8)	1.96	2.88	1.18	2.34	0.78	1.44	-0.79	-0.40
Count words (5)	44.40	11.97	52.50	27.89	63.12	34.47	8.10	10.62
Operations (2)	0.00	0.00	1.55	2.19	0.00	0.00	1.55	-1.55
Comparatives (4)	1.58	3.15	3.13	4.43	4.70	5.43	1.55	1.58
Lesson 1: Zoo (1)	22.42	23.42	28.65	27.91	33.87	30.58	6.23	5.22
Lesson 2: Bakery	18.76	27.17	21.24	29.43	26.24	37.00	2.48	5.00
Lesson 3: Zoo (2)	1.55	1.70	4.18	5.11	2.08	2.56	2.63	-2.10
Lesson 4: Fruit shop	0.62	1.39	0.62	1.39	1.24	1.70	0.00	0.62
Lesson 5: Forest	3.67	3.10	8.92	7.82	11.47	13.52	5.25	2.55
Lesson 6: Playground	3.13	2.82	3.65	3.67	2.08	2.56	0.52	-1.57
Number domain	13.98	21.18	17.84	24.20	20.41	29.00	3.86	2.57
Space domain	2.58	2.78	4.62	6.00	5.15	9.12	2.04	0.53

Table 10
Translation task L2>L1: Tablet only

Lesson	Target word	% correct pretest n=55	% correct pL1 n=54	% correct pL2 n=54	D pretest>pL1	D pL1>pL2
Number domain						
1	One	67.3	70.4	75.9	3.1	5.5
	Two	61.80	72.2	85.2	10.4	13
	Three	52.7	64.8	74.1	12.1	9.3
	More	5.5	18.5	16.7	13	-1.8
	Add	0	0	0	0	0
	Most	0	18.5	25.9	18.5	7.4

2	Four	49.1	66.7	74.1	17.6	7.4
	Five	83.6	85.2	90.7	1.6	5.5
	Fewer	0	1.9	5.6	1.9	3.7
	Take away	1.8	5.6	1.9	3.8	-3.7
	Fewest	0	5.6	9.3	5.6	3.7
3	Big	3.6	11.1	7.4	7.5	-3.7
	Small	3.6	35.2	31.5	31.6	-3.7
	Heavy	1.8	1.9	0	0.1	-1.9
	Light	1.8	1.9	7.4	0.1	5.5
	High	3.6	7.4	3.7	3.8	-3.7
	Low	0	5.6	5.6	5.6	0
Space domain						
4	On	3.6	9.3	3.7	5.7	-5.6
	Above	0	7.4	9.3	7.4	1.9
	Below	0	5.6	5.6	5.6	0
	Next to	0	7.4	9.3	7.4	1.9
	Falling	0	18.5	25.9	18.5	7.4
5	In front of	0	1.9	1.9	1.9	0
	Behind	0	1.9	1.9	1.9	0
	Walking	1.8	7.4	11.1	5.6	3.7
	Running	10.9	63	63	52.1	0
	Jumping	20	70.4	74.1	50.4	3.7
	Flying	9.1	29.6	31.5	20.5	1.9
6	Left	0	1.9	0	1.9	-1.9
	Right	1.8	3.7	0	1.9	-3.7
	Catching	1.8	3.7	3.7	1.9	0
	Throwing	1.8	5.6	3.7	3.8	-1.9
	Sliding	3.6	42.6	53.7	39	11.1
	Climbing	0	24.1	27.8	24.1	3.7

	Pretest		*PL1*		*PL2*		Mean D pre>pL1	Mean D pL1>pL2
	M	SD	M	SD	M	SD		
Movement verbs (9)	5.44	6.69	29.43	24.60	32.72	25.72	23.99	3.29
Measurement words (6)	2.40	1.47	10.52	12.59	9.27	11.24	8.12	-1.25
Prepositions (8)	0.68	1.34	4.89	2.95	3.96	3.77	4.21	-0.93
Count words (5)	62.90	13.63	71.86	36.21	80.00	40.18	8.96	8.14
Operations (2)	0.90	1.27	2.80	3.96	0.95	1.34	1.90	-1.85
Comparatives (4)	1.38	2.75	11.13	8.65	14.38	8.96	9.75	3.25

Lesson 1: Zoo (1)	31.22	32.59	40.73	31.93	46.30	36.33	9.52	5.57
Lesson 2: Bakery	26.90	38.03	33.00	39.78	36.32	42.55	6.10	3.32
Lesson 3: Zoo (2)	2.40	1.47	10.52	12.59	9.27	11.24	8.12	-1.25
Lesson 4: Fruit shop	0.72	1.61	9.64	5.12	10.76	8.80	8.92	1.12
Lesson 5: Forest	6.97	7.92	29.03	31.00	30.58	31.53	22.07	1.55
Lesson 6: Playground	1.50	1.35	13.60	16.42	14.82	21.76	12.10	1.22
Number domain	19.78	29.53	27.79	30.77	30.29	34.32	8.02	2.50
Space domain	3.20	5.40	17.88	21.59	19.19	23.58	14.68	1.31

Table 11
Translation task L2>L1: Robot without iconic gestures

Lesson	Target word	% correct pretest n=57	% correct pL1 n=54	% correct pL2 n=54	D pretest>pL1	D pL1>pL2
Number domain						
1	One	56.1	59.3	70.4	3.2	11.1
	Two	38.60	59.3	70.4	20.7	11.1
	Three	49.1	70.4	81.5	21.3	11.1
	More	7	24.1	24.1	17.1	0
	Add	0	0	0	0	0
	Most	1.8	16.7	20.4	14.9	3.7
2	Four	33.3	55.6	55.6	22.3	0
	Five	82.5	88.9	92.6	6.4	3.7
	Fewer	0	3.7	3.7	3.7	0
	Take away	0	0	0	0	0
	Fewest	0	3.7	5.7	3.7	2
3	Big	5.3	13	9.3	7.7	-3.7
	Small	3.5	25.9	20.4	22.4	-5.5
	Heavy	0	7.4	5.6	7.4	-1.8
	Light	8.8	11.1	14.8	2.3	3.7
	High	3.5	7.4	9.3	3.9	1.9
	Low	0	5.6	9.3	5.6	3.7
Space domain						
4	On	5.3	0	3.7	-5.3	3.7
	Above	0	13	7.4	13	-5.6
	Below	0	1.9	1.9	1.9	0
	Next to	0	11.1	11.1	11.1	0
	Falling	3.5	29.6	22.2	26.1	-7.4
5	In front of	0	0	0	0	0

	Behind	0	1.9	1.9	1.9	0
	Walking	0	7.4	5.6	7.4	-1.8
	Running	14	61.1	55.6	47.1	-5.5
	Jumping	22.8	59.3	57.4	36.5	-1.9
	Flying	19.3	33.3	37	14	3.7
6	Left	1.8	3.7	5.6	1.9	1.9
	Right	3.5	9.3	5.6	5.8	-3.7
	Catching	0	18.5	13	18.5	-5.5
	Throwing	0	3.7	3.7	3.7	0
	Sliding	5.3	53.7	50	48.4	-3.7
	Climbing	0	24.1	27.8	24.1	3.7

	Pretest		*PL1*		*PL2*		Mean D pre>pL1	Mean D pL1>pL2
	M	SD	M	SD	M	SD		
Movement verbs (9)	7.21	9.09	32.30	21.56	30.26	20.91	25.09	-2.04
Measurement words (6)	3.52	3.34	11.73	7.45	11.45	5.28	8.22	-0.28
Prepositions (8)	1.33	2.05	5.11	5.22	4.65	3.56	3.79	-0.46
Count words (5)	51.92	19.27	66.70	34.74	74.10	38.53	14.78	7.40
Operations (2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Comparatives (4)	2.20	3.31	12.05	10.10	13.48	10.28	9.85	1.43
Lesson 1: Zoo (1)	25.43	25.37	38.30	28.45	44.47	33.73	12.87	6.17
Lesson 2: Bakery	23.16	36.17	30.38	40.02	31.52	41.06	7.22	1.14
Lesson 3: Zoo (2)	3.52	3.34	11.73	7.45	11.45	5.28	8.22	-0.28
Lesson 4: Fruit shop	1.76	2.49	11.12	11.77	9.26	8.05	9.36	-1.86
Lesson 5: Forest	9.35	10.62	27.17	28.24	26.25	27.04	17.82	-0.92
Lesson 6: Playground	1.77	2.23	18.83	18.93	17.62	18.20	17.07	-1.22
Number domain	17.03	25.27	26.59	28.45	29.01	31.51	9.56	2.41
Space domain	4.44	7.23	19.51	20.98	18.21	19.94	15.06	-1.30

Table 12
Translation task L2>L1: Robot with iconic gestures

Lesson	Target word	% correct pretest n=60	% correct pL1 n=54	% correct pL2 n=54	D pretest>pL1	D pL1>pL2
Number domain						
1	One	55	61.1	70.4	6.1	9.3

	Two	48.30	61.1	63	12.8	1.9
	Three	46.7	63	72.2	16.3	9.2
	More	6.7	16.7	22.2	10	5.5
	Add	0	0	0	0	0
	Most	0	20.4	20.4	20.4	0
2	Four	28.3	53.7	68.5	25.4	14.8
	Five	70	85.2	88.9	15.2	3.7
	Fewer	0	0	1.9	0	1.9
	Take away	0	5.6	5.6	5.6	0
	Fewest	0	0	0	0	0
3	Big	5	20.4	16.7	15.4	-3.7
	Small	3.3	35.2	24.1	31.9	-11.1
	Heavy	0	1.9	5.7	1.9	3.8
	Light	6.7	7.4	11.1	0.7	3.7
	High	0	9.3	7.4	9.3	-1.9
	Low	0	7.4	11.3	7.4	3.9
Space domain						
4	On	5	1.9	9.3	-3.1	7.4
	Above	0	16.7	18.5	16.7	1.8
	Below	0	0	1.9	0	1.9
	Next to	0	7.4	9.3	7.4	1.9
	Falling	0	27.8	22.2	27.8	-5.6
5	In front of	0	1.9	3.7	1.9	1.8
	Behind	0	1.9	0	1.9	-1.9
	Walking	0	9.3	5.6	9.3	-3.7
	Running	10	57.4	57.4	47.4	0
	Jumping	16.7	61.1	61.1	44.4	0
	Flying	10	38.9	31.5	28.9	-7.4
6	Left	0	1.9	1.9	1.9	0
	Right	1.7	5.6	1.9	3.9	-3.7
	Catching	1.7	5.6	13.2	3.9	7.6
	Throwing	1.7	3.7	3.7	2	0
	Sliding	3.3	50	61.1	46.7	11.1
	Climbing	3.3	16.7	29.6	13.4	12.9

	Pretest		*PL1*		*PL2*		Mean D pre>pL1	Mean D pL1>pL2
	M	SD	M	SD	M	SD		
Movement verbs (9)	5.19	5.75	30.06	22.62	31.71	23.16	*24.87*	*1.66*
Measurement words (6)	2.50	2.94	13.60	12.21	12.72	6.75	*11.10*	*-0.88*

Prepositions (8)	0.84	1.78	4.66	5.42	5.81	6.20	*3.83*	*1.15*
Count words (5)	49.66	15.08	64.82	35.44	72.60	36.27	*15.16*	*7.78*
Operations (2)	0.00	0.00	2.80	3.96	2.80	3.96	*2.80*	*0.00*
Comparatives (4)	1.68	3.35	9.28	10.82	11.13	11.80	*7.60*	*1.85*
Lesson 1: Zoo (1)	26.12	26.42	37.05	27.91	41.37	30.92	*10.93*	*4.32*
Lesson 2: Bakery	19.66	30.69	28.90	38.72	32.98	42.40	*9.24*	*4.08*
Lesson 3: Zoo (2)	2.50	2.94	13.60	12.21	12.72	6.75	*11.10*	*-0.88*
Lesson 4: Fruit shop	1.00	2.24	10.76	11.52	12.24	8.10	*9.76*	*1.48*
Lesson 5: Forest	6.12	7.13	28.42	27.53	26.55	27.70	*22.30*	*-1.87*
Lesson 6: Playground	1.95	1.24	13.92	18.42	18.57	23.39	*11.97*	*4.65*
Number domain	15.88	23.82	26.38	27.76	28.79	30.40	*10.49*	*2.41*
Space domain	3.14	4.79	18.11	20.96	19.52	21.51	*14.96*	*1.42*

Appendix III: Results of in-depth analyses: Translation task L1>L2

D = difference
PL1 = posttest 1
PL2 = posttest 2

Table 13
Translation task L1>L2: Control condition

Lesson	Target word	% correct pL1 n=31	% correct pL2 n=31	D pL1>pL2
Number domain				
1	One	65.6	65.6	0
	Two	50	71.9	21.9
	Three	43.8	56.3	12.5
	More	6.3	0	-6.3
	Add	0	0	0
	Most	6.3	0	-6.3
2	Four	43.8	62.5	18.7
	Five	37.5	46.9	9.4
	Fewer	0	0	0
	Take away	0	0	0
	Fewest	3.1	0	-3.1
3	Big	0	0	0
	Small	0	3.1	3.1
	Heavy	0	0	0
	Light	9.4	0	-9.4
	High	0	0	0
	Low	0	0	0
Space domain				
4	On	0	0	0
	Above	0	0	0
	Below	3.1	0	-3.1
	Next to	0	0	0
	Falling	6.3	3.1	-3.2
5	In front of	0	0	0
	Behind	0	0	0
	Walking	0	0	0
	Running	0	6.3	6.3
	Jumping	6.3	12.5	6.2
	Flying	6.3	9.4	3.1
6	Left	0	0	0
	Right	0	3.1	3.1

Catching	6.3	0	-6.3
Throwing	0	0	0
Sliding	3.1	0	-3.1
Climbing	12.5	3.1	-9.4

	PL1		*PL2*		Mean D pL1>pL2
	M	SD	M	SD	
Movement verbs (9)	4.53	4.18	3.82	4.64	-0.71
Measurement words (6)	1.57	3.84	0.52	1.27	-1.05
Prepositions (8)	0.39	1.10	0.39	1.10	0.00
Count words (5)	48.14	22.79	60.64	33.19	12.50
Operations (2)	0.00	0.00	0.00	0.00	0.00
Comparatives (4)	3.93	3.02	0.00	0.00	-3.93
Lesson 1: Zoo (1)	28.67	27.82	32.30	35.73	3.63
Lesson 2: Bakery	16.88	21.85	21.88	30.46	5.00
Lesson 3: Zoo (2)	1.57	3.84	0.52	1.27	-1.05
Lesson 4: Fruit shop	1.88	2.81	0.62	1.39	-1.26
Lesson 5: Forest	2.10	3.25	4.70	5.51	2.60
Lesson 6: Playground	3.65	5.01	1.03	1.60	-2.62
Number domain	15.64	22.46	18.02	28.77	2.38
Space domain	2.58	3.72	2.21	3.79	-0.38

Table 14
Translation task L1>L2: Tablet only

Lesson	Target word	% correct pL1 n=54	% correct pL2 n=54	D pL1>pL2
Number domain				
1	One	74.1	81.5	7.4
	Two	75.9	81.5	5.6
	Three	68.5	68.5	0
	More	16.7	16.7	0
	Add	3.7	0	-3.7
	Most	14.8	18.5	3.7
2	Four	74.1	81.5	7.4
	Five	72.2	63	-9.2
	Fewer	1.9	1.9	0
	Take away	1.9	1.9	0
	Fewest	0	0	0

3	Big	7.4	7.4	0
	Small	24.1	25.9	1.8
	Heavy	1.9	0	-1.9
	Light	18.5	14.8	-3.7
	High	1.9	3.7	1.8
	Low	7.5	0	-7.5
Space domain				
4	On	1.9	0	-1.9
	Above	0	0	0
	Below	1.9	3.7	1.8
	Next to	3.7	1.9	-1.8
	Falling	29.6	22.2	-7.4
5	In front of	0	0	0
	Behind	1.9	3.7	1.8
	Walking	7.4	5.6	-1.8
	Running	18.5	25.9	7.4
	Jumping	33.3	40.7	7.4
	Flying	35.2	33.3	-1.9
6	Left	0	1.9	1.9
	Right	5.6	7.4	1.8
	Catching	3.7	1.9	-1.8
	Throwing	1.9	1.9	0
	Sliding	22.2	22.2	0
	Climbing	15.1	24.1	9

	PL1		*PL2*		Mean D pL1>pL2
	M	SD	M	SD	
Movement verbs (9)	18.54	12.57	19.76	13.82	1.21
Measurement words (6)	10.22	9.11	8.63	10.11	-1.58
Prepositions (8)	1.88	1.99	2.33	2.57	0.45
Count words (5)	72.96	34.94	75.20	34.85	2.24
Operations (2)	2.80	1.27	0.95	1.34	-1.85
Comparatives (4)	8.35	8.61	9.28	9.67	0.93
Lesson 1: Zoo (1)	42.28	33.85	44.45	36.72	2.17
Lesson 2: Bakery	30.02	39.39	29.66	39.43	-0.36
Lesson 3: Zoo (2)	10.22	9.11	8.63	10.11	-1.58
Lesson 4: Fruit shop	7.42	12.47	5.56	9.43	-1.86
Lesson 5: Forest	16.05	15.51	18.20	17.28	2.15
Lesson 6: Playground	8.08	8.69	9.90	10.50	1.82

Number domain	27.36	31.11	27.46	32.94	0.10
Space domain	10.70	12.42	11.55	13.37	0.85

Table 15
Translation task L1>L2: Robot without iconic gestures

Lesson	Target word	% correct pL1 n=54	% correct pL2 n=54	D pL1>pL2
Number domain				
1	One	72.2	77.8	5.6
	Two	81.5	66.7	-14.8
	Three	72.2	70.4	-1.8
	More	14.8	20.4	5.6
	Add	3.7	0	-3.7
	Most	13	20.4	7.4
2	Four	68.5	63	-5.5
	Five	70.4	61.1	-9.3
	Fewer	0	1.9	1.9
	Take away	0	0	0
	Fewest	0	1.9	1.9
3	Big	7.4	13	5.6
	Small	13	11.1	-1.9
	Heavy	0	0	0
	Light	3.7	11.1	7.4
	High	11.1	9.4	-1.7
	Low	3.7	1.9	-1.8
Space domain				
4	On	0	1.9	1.9
	Above	0	1.9	1.9
	Below	0	3.7	3.7
	Next to	3.7	3.7	0
	Falling	31.5	33.3	1.8
5	In front of	1.9	0	-1.9
	Behind	0	0	0
	Walking	5.6	5.6	0
	Running	29.6	25.9	-3.7
	Jumping	38.9	37	-1.9
	Flying	29.6	31.5	1.9
6	Left	3.7	3.7	0
	Right	9.3	5.6	-3.7
	Catching	9.3	7.4	-1.9
	Throwing	0	0	0

Sliding	31.5	31.5	0
Climbing	27.8	22.2	-5.6

	PL1		*PL2*		Mean D pL1>pL2
	M	SD	M	SD	
Movement verbs (9)	22.64	13.81	21.60	13.75	-1.04
Measurement words (6)	6.48	4.94	7.75	5.42	1.27
Prepositions (8)	2.33	3.26	2.56	1.97	0.24
Count words (5)	72.96	37.47	67.80	30.96	-5.16
Operations (2)	1.85	2.62	0.00	0.00	-1.85
Comparatives (4)	6.95	8.06	11.15	10.68	4.20
Lesson 1: Zoo (1)	42.90	35.85	42.62	32.84	-0.28
Lesson 2: Bakery	27.78	38.05	25.58	33.31	-2.20
Lesson 3: Zoo (2)	6.48	4.94	7.75	5.42	1.27
Lesson 4: Fruit shop	7.04	13.77	8.90	13.67	1.86
Lesson 5: Forest	17.60	16.98	16.67	16.71	-0.93
Lesson 6: Playground	13.60	12.98	11.73	12.32	-1.87
Number domain	25.60	31.97	25.30	29.18	-0.30
Space domain	13.08	14.47	12.64	13.86	-0.44

Table 16
Translation task L1>L2: Robot with iconic gestures

Lesson	Target word	% correct pL1 n=54	% correct pL2 n=54	D pL1>pL2
Number domain				
1	One	70.4	77.8	7.4
	Two	63	64.8	1.8
	Three	64.8	63	-1.8
	More	7.4	13	5.6
	Add	1.9	0	-1.9
	Most	22.2	14.8	-7.4
2	Four	63	75.9	12.9
	Five	63	61.1	-1.9
	Fewer	1.9	0	-1.9
	Take away	3.7	3.7	0

	Fewest	1.9	1.9	0
3	Big	13	14.8	1.8
	Small	22.2	29.6	7.4
	Heavy	3.7	1.9	-1.8
	Light	7.4	13	5.6
	High	5.6	5.6	0
	Low	5.6	1.9	-3.7
Space domain				
4	On	1.9	1.9	0
	Above	1.9	0	-1.9
	Below	1.9	3.7	1.8
	Next to	1.9	11.1	9.2
	Falling	18.5	18.5	0
5	In front of	1.9	0	-1.9
	Behind	0	0	0
	Walking	3.7	3.7	0
	Running	25.9	24.1	-1.8
	Jumping	35.2	42.6	7.4
	Flying	24.1	35.2	11.1
6	Left	0	1.9	1.9
	Right	0	3.7	3.7
	Catching	5.6	11.1	5.5
	Throwing	0	0	0
	Sliding	38.9	37	-1.9
	Climbing	24.1	18.5	-5.6

	PL1		*PL2*		Mean D pL1>pL2
	M	SD	M	SD	
Movement verbs (9)	19.56	13.83	21.19	14.93	1.63
Measurement words (6)	9.58	6.96	11.13	10.57	1.55
Prepositions (8)	1.19	0.98	2.79	3.70	1.60
Count words (5)	64.84	32.36	68.52	31.30	3.68
Operations (2)	2.80	1.27	1.85	2.62	-0.95
Comparatives (4)	8.35	9.59	7.43	7.55	-0.93
Lesson 1: Zoo (1)	38.28	31.25	38.90	33.26	0.62
Lesson 2: Bakery	26.70	33.15	28.52	36.89	1.82
Lesson 3: Zoo (2)	9.58	6.96	11.13	10.57	1.55
Lesson 4: Fruit shop	5.22	7.42	7.04	7.66	1.82

Lesson 5: Forest	15.13	15.06	17.60	18.92	2.47
Lesson 6: Playground	11.43	16.38	12.03	14.02	0.60
Number domain	24.75	27.40	26.05	29.46	1.30
Space domain	10.91	13.62	12.53	14.39	1.62

Appendix IV: Results of in-depth analyses: Comprehension Task

D = difference
PL1 = posttest 1
PL2 = posttest 2

Table 17
Comprehension task: Control condition

	Posttest 1					Posttest 2					Difference posttest 1 > posttest 2			
Word	*n*	0 trials correct	1 trial correct	2 trials correct	3 trials correct	***n***	0 trials correct	1 trial correct	2 trials correct	3 trials correct	0 trials correct	1 trial correct	2 trials correct	3 trials correct
Two	31	9.7	6.5	29	54.8	30	10	13.3	16.7	60	-21	-6.5	-12.3	5.2
Add	30	20	53.3	20	6.7	29	24.1	37.9	37.9	0	-5.9	-43.3	17.9	-6.7
Most	31	9.7	16.1	41.9	32.3	30	6.7	10	36.7	46.7	-24.3	0.6	-5.2	14.4
Four	31	6.5	19.4	22.6	51.6	30	6.7	16.7	40	36.7	-24.3	-19.4	17.4	-14.9
Take away	29	13.8	65.5	13.8	6.9	29	37.9	37.9	20.7	3.4	8.9	-36.9	6.9	-3.5
Fewest	29	34.5	48.3	13.8	3.4	28	57.1	28.6	14.3	0	28.1	-48.3	0.5	-3.4
Small	30	3.3	13.3	30	53.3	30	10	13.3	23.3	53.3	-20	21.2	-6.7	0
Heavy	29	17.2	51.7	24.1	6.9	29	17.2	34.5	44.8	3.4	-11.8	-51.7	20.7	-3.5
Low	30	33.3	36.7	20	10	29	24.1	48.3	20.7	6.9	-5.9	-10	0.7	-3.1
On	30	13.3	26.7	53.3	6.7	30	13.3	26.7	46.7	13.3	-16.7	-26.7	-6.6	6.6
Below	29	20.7	58.6	13.8	6.9	30	26.7	30	40	3.3	-2.3	-25.3	26.2	-3.6
Next to	30	23.3	46.7	10	20	30	30	33.3	26.7	10	0	-46.7	16.7	-10
In front of	30	63.3	23.3	13.3	0	30	46.7	46.7	6.7	0	16.7	4.3	-6.6	0
Behind	30	33.3	40	16.7	10	29	27.6	27.6	37.9	6.9	-2.4	-40	21.2	-3.1
Jumping	30	6.7	16.7	20	56.7	30	6.7	23.3	16.7	53.3	-23.3	-16.7	-3.3	-3.4
Catching	30	26.7	36.7	26.7	10	30	20	43.3	26.7	10	-10	-36.7	0	0
Sliding	30	23.3	23.3	26.7	26.7	30	10	26.7	13.3	50	-20	-10	-13.4	23.3

Climbing	29	20.7	31	34.5	13.8	30	16.7	13.3	36.7	33	-12.3	-31	2.2	19.2		

	Posttest 1								**Posttest 2**							
	0 trials correct		*1 trial correct*		*2 trials correct*		*3 trials correct*		*0 trials correct*		*1 trial correct*		*2 trials correct*		*3 trials correct*	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Movement verbs (9)	19.35	8.78	26.93	8.75	26.98	5.93	26.80	21.18	13.35	6.08	26.65	12.47	23.35	10.56	36.58	19.82
Measurement words	17.93	15.01	33.90	19.35	24.70	5.03	23.40	25.94	17.10	7.05	32.03	17.63	29.60	13.23	21.20	27.85
Prepositions (8)	30.78	19.54	39.06	14.51	21.42	17.98	8.72	7.29	28.86	11.92	32.86	8.15	31.60	15.67	6.70	5.27
Count words (5)	8.10	2.26	12.95	9.12	25.80	4.53	53.20	2.26	8.35	2.33	15.00	2.40	28.35	16.48	48.35	16.48
Operations (2)	16.90	4.38	59.40	8.63	16.90	4.38	6.80	0.14	31.00	9.76	37.90	0.00	29.30	12.16	1.70	2.40
Comparatives (4)	22.10	17.54	32.20	22.77	27.85	19.87	17.85	20.44	31.90	35.64	19.30	13.15	25.50	15.84	23.35	33.02
Lesson 1: Zoo (1)	13.13	5.95	25.30	24.72	30.30	11.01	31.27	24.07	13.60	9.24	20.40	15.24	30.43	11.91	35.57	31.51
Lesson 2: Bakery	18.27	14.52	44.40	23.30	16.73	5.08	20.63	26.87	33.90	25.44	27.73	10.63	25.00	13.38	13.37	20.28
Lesson 3: Zoo (2)	17.93	15.01	33.90	19.35	24.70	5.03	23.40	25.94	17.10	7.05	32.03	17.63	29.60	13.23	21.20	27.85
Lesson 4: Fruit shop	19.10	5.19	44.00	16.12	25.70	23.98	11.20	7.62	23.33	8.84	30.00	3.30	37.80	10.18	8.87	5.10
Lesson 5: Forest	34.43	28.32	26.67	12.01	16.67	3.35	22.23	30.26	27.00	20.01	32.53	12.46	20.43	15.93	20.07	28.99
Lesson 6: Playground	23.57	3.01	30.33	6.72	29.30	4.50	16.83	8.75	15.57	5.10	27.77	15.03	25.57	11.74	31.00	20.07
Number domain	16.44	11.14	34.53	21.23	23.91	8.83	25.10	22.72	21.53	16.85	26.72	13.78	28.34	11.42	23.38	25.30
Space domain	25.70	16.00	33.67	13.23	23.89	13.54	16.76	16.90	21.97	12.32	30.10	10.11	27.93	13.55	19.98	20.23

	Mean difference posttest 1 > posttest 2			
	0 trials correct	1 trial correct	2 trials correct	3 trials correct
Movement verbs (4)	-6.00	-0.28	-3.63	9.78
Measurement words (3)	-0.83	-1.87	4.90	-2.20
Prepositions (5)	-1.92	-6.20	10.18	-2.02
Count words (2)	0.25	2.05	2.55	-4.85
Operations (2)	14.10	-21.50	12.40	-5.10
Comparatives (2)	9.80	-12.90	-2.35	5.50
Lesson 1: Zoo (1)	0.47	-4.90	0.13	4.30
Lesson 2: Bakery	15.63	-16.67	8.27	-7.27
Lesson 3: Zoo (2)	-0.83	-1.87	4.90	-2.20
Lesson 4: Fruit shop	4.23	-14.00	12.10	-2.33
Lesson 5: Forest	-7.43	5.87	3.77	-2.17
Lesson 6: Playground	-8.00	-2.57	-3.73	14.17
Number domain	5.09	-7.81	4.43	-1.72
Space domain	-3.73	-3.57	4.04	3.22

Table 18
Comprehension task: Tablet only

	Posttest 1					Posttest 2					Difference posttest 1 > posttest 2			
Word	*n*	0 trials correct	1 trial correct	2 trials correct	3 trials correct	***n***	0 trials correct	1 trial correct	2 trials correct	3 trials correct	0 trials correct	1 trial correct	2 trials correct	3 trials correct
Two	51	5.9	7.8	15.7	70.6	53	3.8	15.1	15.1	66	-47.2	-7.8	-0.6	-4.6
Add	50	28	34	32	6	51	23.5	37.3	25.5	13.7	-26.5	-18.6	-6.5	7.7
Most	50	6	12	26	56	52	9.6	15.4	25	50	-40.4	9.2	-1	-6
Four	50	2	14	28	56	52	3.8	21.2	11.5	63.5	-46.2	-14	-16.5	7.5
Take away	51	33.3	52.9	13.7	0	52	25	42.3	21.2	11.5	-26	-14.4	7.5	11.5
Fewest	48	45.8	35.4	8.3	10.4	52	42.3	38.5	5.8	13.5	-5.7	-35.4	-2.5	3.1
Small	52	0	7.7	25	67.3	51	5.9	9.8	5.9	78.4	-46.1	35.4	-19.1	11.1
Heavy	48	20.8	41.7	22.9	14.6	51	19.6	43.1	27.5	9.8	-28.4	-41.7	4.6	-4.8
Low	51	27.5	41.2	17.6	13.7	52	30.8	34.6	15.4	19.2	-20.2	-20.4	-2.2	5.5
On	52	7.7	44.2	34.6	13.5	53	15.1	20.8	47.2	17	-36.9	-44.2	12.6	3.5
Below	52	7.7	38.5	51.9	1.9	53	11.3	47.2	30.2	11.3	-40.7	-5.2	-21.7	9.4
Next to	50	16	32	30	22	51	19.6	33.3	23.5	23.5	-30.4	-32	-6.5	1.5
In front of	51	43.1	39.2	13.7	3.9	52	46.2	26.9	21.2	5.8	-4.8	8.8	7.5	1.9
Behind	50	30	42	22	6	50	24	48	20	8	-26	-42	-2	2
Jumping	51	2	13.7	9.8	74.5	52	1.9	15.4	9.6	73.1	-49.1	-13.7	-0.2	-1.4
Catching	51	15.7	31.4	37.3	15.7	52	17.3	36.5	32.7	13.5	-33.7	-31.4	-4.6	-2.2
Sliding	52	11.5	5.8	17.3	65.4	53	3.8	17	7.5	71.7	-48.2	10.5	-9.8	6.3
Climbing	51	5.9	33.3	25.5	35.3	49	6.1	16.3	26.5	51	-44.9	-33.3	1	15.7

	Posttest 1				Posttest 2			
	0 trials correct	1 trial correct	2 trials correct	3 trials correct	0 trials correct	1 trial correct	2 trials correct	3 trials correct

	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Movement verbs (9)	8.78	6.04	21.05	13.46	22.48	11.78	47.73	27.14	7.28	6.90	21.30	10.15	19.08	12.44	52.33	27.79
Measurement words (6)	16.10	14.34	30.20	19.49	21.83	3.81	31.87	30.69	18.77	12.47	29.17	17.30	16.27	10.83	35.80	37.19
Prepositions (8)	20.90	15.40	39.18	4.61	30.44	14.40	9.46	8.27	23.24	13.69	35.24	12.12	28.42	11.21	13.12	7.17
Count words (5)	3.95	2.76	10.90	4.38	21.85	8.70	63.30	10.32	3.80	0.00	18.15	4.31	13.30	2.55	64.75	1.77
Operations (2)	30.65	3.75	43.45	13.36	22.85	12.94	3.00	4.24	24.25	1.06	39.80	3.54	23.35	3.04	12.60	1.56
Comparatives (4)	25.90	28.14	23.70	16.55	17.15	12.52	33.20	32.24	25.95	23.12	26.95	16.33	15.40	13.58	31.75	25.81
Lesson 1: Zoo (1)	13.30	12.73	17.93	14.07	24.57	8.24	44.20	33.88	12.30	10.12	22.60	12.73	21.87	5.87	43.23	26.80
Lesson 2: Bakery	27.03	22.56	34.10	19.48	16.67	10.18	22.13	29.79	23.70	19.28	34.00	11.25	12.83	7.79	29.50	29.46
Lesson 3: Zoo (2)	16.10	14.34	30.20	19.49	21.83	3.81	31.87	30.69	18.77	12.47	29.17	17.30	16.27	10.83	35.80	37.19
Lesson 4: Fruit shop	10.47	4.79	38.23	6.10	38.83	11.55	12.47	10.09	15.33	4.15	33.77	13.21	33.63	12.22	17.27	6.10
Lesson 5: Forest	25.03	21.00	31.63	15.59	15.17	6.23	28.13	40.17	24.03	22.15	30.10	16.53	16.93	6.38	28.97	38.24
Lesson 6: Playground	11.03	4.92	23.50	15.36	26.70	10.05	38.80	25.03	9.07	7.22	23.27	11.47	22.23	13.13	45.40	29.50
Number domain	18.81	16.08	27.41	17.11	21.02	7.66	32.73	28.91	18.26	13.49	28.59	13.10	16.99	8.29	36.18	27.89
Space domain	15.51	13.15	31.12	13.04	26.90	13.17	26.47	26.78	16.14	13.50	29.04	12.89	24.27	12.05	30.54	27.24

Mean difference posttest 1 > posttest 2				
	0 trials correct	*1 trial correct*	*2 trials correct*	*3 trials correct*
Movement verbs (4)	-1.50	0.25	-3.40	4.60

Measurement words (3)	2.67	-1.03	-5.57	3.93
Prepositions (5)	2.34	-3.94	-2.02	3.66
Count words (2)	-0.15	7.25	-8.55	1.45
Operations (2)	-6.40	-3.65	0.50	9.60
Comparatives (2)	0.05	3.25	-1.75	-1.45
Lesson 1: Zoo (1)	-1.00	4.67	-2.70	-0.97
Lesson 2: Bakery	-3.33	-0.10	-3.83	7.37
Lesson 3: Zoo (2)	2.67	-1.03	-5.57	3.93
Lesson 4: Fruit shop	4.87	-4.47	-5.20	4.80
Lesson 5: Forest	-1.00	-1.53	1.77	0.83
Lesson 6: Playground	-1.97	-0.23	-4.47	6.60
Number domain	-0.56	1.18	-4.03	3.44
Space domain	0.63	-2.08	-2.63	4.08

Table 19
Comprehension task: Robot without iconic gestures

	Posttest 1	**Posttest 2**	**Difference posttest 1 > posttest 2**

Word	*n*	0 trials correct	1 trial correct	2 trials correct	3 trials correct	***n***	0 trials correct	1 trial correct	2 trials correct	3 trials correct	0 trials correct	1 trial correct	2 trials correct	3 trials correct
Two	54	22.2	11.1	11.1	55.6	52	3.8	7.7	19.2	69.2	-50.2	-11.1	8.1	13.6
Add	52	26.9	32.7	34.6	5.8	52	15.4	46.2	34.6	3.8	-36.6	-23.3	0	-2
Most	53	1.9	20.8	28.3	49.1	53	5.7	9.4	28.3	56.6	-47.3	-7.6	0	7.5
Four	53	5.7	15.1	28.3	50.9	53	5.7	13.2	20.8	60.4	-47.3	-15.1	-7.5	9.5
Take away	52	25	57.7	15.4	1.9	52	28.8	44.2	21.2	5.8	-23.2	-25.6	5.8	3.9
Fewest	53	39.6	30.2	22.6	7.5	53	47.2	32.1	15.1	5.7	-5.8	-30.2	-7.5	-1.8
Small	54	3.7	13	18.5	64.8	54	3.7	3.7	27.8	64.8	-50.3	25	9.3	0
Heavy	52	11.5	26.9	53.8	7.7	50	10	38	34	18	-42	-26.9	-19.8	10.3
Low	51	27.5	43.1	7.8	21.6	51	33.3	25.5	15.7	25.5	-17.7	-20.5	7.9	3.9
On	52	9.6	28.8	48.1	13.5	53	7.5	22.6	49.1	20.8	-44.5	-28.8	1	7.3
Below	53	22.6	34	34	9.4	53	32.1	30.2	32.1	5.7	-20.9	3.7	-1.9	-3.7
Next to	53	22.6	30.2	24.5	22.6	53	22.6	37.7	17	22.6	-30.4	-30.2	-7.5	0
In front of	52	50	30.8	19.2	0	54	48.1	40.7	11.1	0	-3.9	29.6	-8.1	0
Behind	52	25	42.3	28.8	3.8	53	24.5	60.4	15.1	0	-27.5	-42.3	-13.7	-3.8
Jumping	54	1.9	16.7	14.8	66.7	54	3.7	22.2	7.4	66.7	-50.3	-16.7	-7.4	0
Catching	51	11.8	37.3	17.6	33.3	53	18.9	26.4	22.6	32.1	-32.1	-37.3	5	-1.2
Sliding	54	7.4	7.4	16.7	68.5	53	15.1	3.8	15.1	66	-38.9	20.9	-1.6	-2.5
Climbing	53	11.3	11.3	35.8	41.5	53	7.5	28.3	18.9	45.3	-45.5	-11.3	-16.9	3.8

	Posttest 1								**Posttest 2**							
	0 trials correct		1 trial correct		2 trials correct		3 trials correct		0 trials correct		1 trial correct		2 trials correct		3 trials correct	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Movement verbs (9)	8.10	4.58	18.18	13.31	21.23	9.79	52.50	17.77	11.30	6.94	20.18	11.21	16.00	6.50	52.53	16.85

Measurement words (6)	14.23	12.13	27.67	15.06	26.70	24.07	31.37	29.78	15.67	15.59	22.40	17.36	25.83	9.31	36.10	25.14
Prepositions (8)	25.96	14.74	33.22	5.42	30.92	11.04	9.86	8.80	26.96	14.81	38.32	14.20	24.88	15.70	9.82	11.11
Count words (5)	13.95	11.67	13.10	2.83	19.70	12.16	53.25	3.32	4.75	1.34	10.45	3.89	20.00	1.13	64.80	6.22
Operations (2)	25.95	1.34	45.20	17.68	25.00	13.58	3.85	2.76	22.10	9.48	45.20	1.41	27.90	9.48	4.80	1.41
Comparatives (4)	20.75	26.66	25.50	6.65	25.45	4.03	28.30	29.42	26.45	29.34	20.75	16.05	21.70	9.33	31.15	35.99
Lesson 1: Zoo (1)	17.00	13.29	21.53	10.82	24.67	12.16	36.83	27.07	8.30	6.22	21.10	21.75	27.37	7.74	43.20	34.70
Lesson 2: Bakery	23.43	17.00	34.33	21.60	22.10	6.46	20.10	26.82	27.23	20.79	29.83	15.62	19.03	3.41	23.97	31.55
Lesson 3: Zoo (2)	14.23	12.13	27.67	15.06	26.70	24.07	31.37	29.78	15.67	15.59	22.40	17.36	25.83	9.31	36.10	25.14
Lesson 4: Fruit shop	18.27	7.51	31.00	2.69	35.53	11.87	15.17	6.76	20.73	12.41	30.17	7.55	32.73	16.06	16.37	9.28
Lesson 5: Forest	25.63	24.06	29.93	12.82	20.93	7.16	23.50	37.46	25.43	22.21	41.10	19.10	11.20	3.85	22.23	38.51
Lesson 6: Playground	10.17	2.41	18.67	16.25	23.37	10.78	47.77	18.42	13.83	5.80	19.50	13.63	18.87	3.75	47.80	17.09
Number domain	18.22	13.04	27.84	15.28	24.49	14.01	29.43	25.28	17.07	15.71	24.44	16.47	24.08	7.37	34.42	27.91
Space domain	18.02	14.32	26.53	12.00	26.61	11.09	28.81	25.73	20.00	13.99	30.26	15.47	20.93	12.69	28.80	25.98

Mean difference posttest 1 > posttest 2				
	0 trials correct	*1 trial correct*	*2 trials correct*	*3 trials correct*
Movement verbs (4)	3.20	2.00	-5.23	0.03
Measurement words (3)	1.43	-5.27	-0.87	4.73
Prepositions (5)	1.00	5.10	-6.04	-0.04
Count words (2)	-9.20	-2.65	0.30	11.55

Operations (2)	-3.85	0.00	2.90	0.95
Comparatives (2)	5.70	-4.75	-3.75	2.85
Lesson 1: Zoo (1)	-8.70	-0.43	2.70	6.37
Lesson 2: Bakery	3.80	-4.50	-3.07	3.87
Lesson 3: Zoo (2)	1.43	-5.27	-0.87	4.73
Lesson 4: Fruit shop	2.47	-0.83	-2.80	1.20
Lesson 5: Forest	-0.20	11.17	-9.73	-1.27
Lesson 6: Playground	3.67	0.83	-4.50	0.03
Number domain	-1.16	-3.40	-0.41	4.99
Space domain	1.98	3.72	-5.68	-0.01

Table 20
Comprehension task: Robot with iconic gestures

	Posttest 1					**Posttest 2**					**Difference posttest 1 > posttest 2**			
Word	*n*	0 trials correct	1 trial correct	2 trials correct	3 trials correct	***n***	0 trials correct	1 trial correct	2 trials correct	3 trials correct	0 trials correct	1 trial correct	2 trials correct	3 trials correct
Two	53	11.3	7.5	20.8	60.4	53	5.7	9.4	13.2	71.7	-47.3	-7.5	-7.6	11.3
Add	49	16.3	49	30.6	4.1	54	16.7	33.3	38.9	11.1	-32.3	-41.5	8.3	7
Most	53	5.7	13.2	37.7	43.4	53	3.8	7.5	28.3	60.4	-49.2	-2.1	-9.4	17

Four	53	7.5	7.5	30.2	54.7	54	3.7	11.1	18.5	66.7	-49.3	-7.5	-11.7	12
Take away	53	28.3	49.1	18.9	3.8	54	27.8	48.1	13	11.1	-25.2	-15.8	-5.9	7.3
Fewest	51	37.3	37.3	15.7	9.8	54	46.3	33.3	11.1	9.3	-4.7	-37.3	-4.6	-0.5
Small	53	5.7	0	15.1	79.2	54	1.9	9.3	24.1	64.8	-51.1	31.5	9	-14.4
Heavy	50	12	34	46	8	54	16.7	31.5	37	14.8	-33.3	-34	-9	6.8
Low	52	30.8	38.5	21.2	9.6	54	29.6	42.6	13	14.8	-22.4	-7	-8.2	5.2
On	52	7.7	36.5	38.5	17.3	54	16.7	31.5	33.3	18.5	-35.3	-36.5	-5.2	1.2
Below	51	19.6	45.1	33.3	2	53	24.5	30.2	34	11.3	-26.5	-13.6	0.7	9.3
Next to	50	30	38	22	10	54	22.2	31.5	25.9	20.4	-27.8	-38	3.9	10.4
In front of	53	41.5	37.7	17	3.8	54	50	42.6	7.4	0	-3	4.9	-9.6	-3.8
Behind	52	38.5	42.3	17.3	1.9	54	29.6	42.6	22.2	5.6	-22.4	-42.3	4.9	3.7
Jumping	53	3.8	11.3	17	67.9	53	3.8	7.5	17	71.7	-49.2	-11.3	0	3.8
Catching	53	11.3	28.3	32.1	28.3	54	14.8	29.6	25.9	29.6	-38.2	-28.3	-6.2	1.3
Sliding	53	3.8	1.9	20.8	73.6	53	3.8	9.4	17	69.8	-49.2	20.3	-3.8	-3.8
Climbing	52	15.4	17.3	28.8	38.5	54	11.1	22.2	25.9	40.7	-40.9	-17.3	-2.9	2.2

	Posttest 1								**Posttest 2**							
	0 trials correct		1 trial correct		2 trials correct		3 trials correct		0 trials correct		1 trial correct		2 trials correct		3 trials correct	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Movement verbs (9)	8.58	5.76	14.70	11.06	24.68	6.98	52.08	22.09	8.38	5.49	17.18	10.55	21.45	5.14	52.95	21.06
Measurement words (6)	16.17	13.06	24.17	21.05	27.43	16.37	32.27	40.65	16.07	13.86	27.80	16.96	24.70	12.01	31.47	28.87
Prepositions (8)	27.46	13.94	39.92	3.63	25.62	9.77	7.00	6.64	28.60	12.83	35.68	6.34	24.56	10.81	11.16	8.58

Count words (5)	9.40	2.69	7.50	0.00	25.50	6.65	57.55	4.03	4.70	1.41	10.25	1.20	15.85	3.75	69.20	3.54
Operations (2)	22.30	8.49	49.05	0.07	24.75	8.27	3.95	0.21	22.25	7.85	40.70	10.47	25.95	18.31	11.10	0.00
Comparatives (4)	21.50	22.34	25.25	17.04	26.70	15.56	26.60	23.76	25.05	30.05	20.40	18.24	19.70	12.16	34.85	36.13
Lesson 1: Zoo (1)	11.10	5.30	23.23	22.50	29.70	8.49	35.97	28.88	8.73	6.96	16.73	14.38	26.80	12.92	47.73	32.22
Lesson 2: Bakery	24.37	15.28	31.30	21.44	21.60	7.62	22.77	27.82	25.93	21.36	30.83	18.62	14.20	3.84	29.03	32.63
Lesson 3: Zoo (2)	16.17	13.06	24.17	21.05	27.43	16.37	32.27	40.65	16.07	13.86	27.80	16.96	24.70	12.01	31.47	28.87
Lesson 4: Fruit shop	19.10	11.16	39.87	4.59	31.27	8.44	9.77	7.65	21.13	4.01	31.07	0.75	31.07	4.49	16.73	4.80
Lesson 5: Forest	27.93	20.95	30.43	16.73	17.10	0.17	24.53	37.57	27.80	23.15	30.90	20.26	15.53	7.51	25.77	39.88
Lesson 6: Playground	10.17	5.88	15.83	13.26	27.23	5.81	46.80	23.76	9.90	5.60	20.40	10.22	22.93	5.14	46.70	20.76
Number domain	17.21	11.90	26.23	19.15	26.24	10.61	30.33	29.15	16.91	15.17	25.12	15.86	21.90	10.75	36.08	28.49
Space domain	19.07	14.45	28.71	15.14	25.20	8.14	27.03	27.74	19.61	14.40	27.46	12.53	23.18	8.43	29.73	26.24

	Mean difference posttest 1 > posttest 2			
	0 trials correct	*1 trial correct*	*2 trials correct*	*3 trials correct*
Movement verbs (4)	-0.20	2.48	-3.23	0.88
Measurement words (3)	-0.10	3.63	-2.73	-0.80
Prepositions (5)	1.14	-4.24	-1.06	4.16
Count words (2)	-4.70	2.75	-9.65	11.65
Operations (2)	-0.05	-8.35	1.20	7.15
Comparatives (2)	3.55	-4.85	-7.00	8.25

Lesson 1: Zoo (1)	-2.37	-6.50	-2.90	11.77
Lesson 2: Bakery	1.57	-0.47	-7.40	6.27
Lesson 3: Zoo (2)	-0.10	3.63	-2.73	-0.80
Lesson 4: Fruit shop	2.03	-8.80	-0.20	6.97
Lesson 5: Forest	-0.13	0.47	-1.57	1.23
Lesson 6: Playground	-0.27	4.57	-4.30	-0.10
Number domain	-0.30	-1.11	-4.34	5.74
Space domain	0.54	-1.26	-2.02	2.70

Appendix V: Results of in-depth analyses: Test at the end of each lesson

Table 21
Test at the end of each lesson: Intervention conditions together

Lesson	Target word	0 trials correct	1 trial correct	2 trials correct
Number domain				
1	One	14.6	22	63.4
	Two	14	24.4	61.6
	Three	25	20.1	54.9
	More	29.9	33.5	36.6
	Add	47.9	45.4	6.7
	Most	28.7	27.4	43.9
2	Four	11.7	21.6	66.7
	Five	0.6	3.1	96.3
	Fewer	47.5	29.6	22.8
	Take away	50.3	38	11.7
	Fewest	57.1	27.6	15.3
3	Big	31.1	30.4	38.5
	Small	10.5	20.4	69.1
	Heavy	36.4	35.8	27.8
	Light	27	39	34
	High	29	38.9	32.1
	Low	63	24.7	12.3
Space domain				
4	On	68.3	25.5	6.2
	Above	42.9	36	21.1
	Below	18	46	36
	Next to	33.5	48.4	18
	Falling	18.1	29.4	52.5
5	In front of	46.5	42.8	10.7
	Behind	55.6	39.4	5
	Walking	73.6	18.2	8.2
	Running	28.7	28.1	43.1
	Jumping	34.6	37.7	27.7
	Flying	21.7	31.1	47.2
6	Left	39.9	43.6	16.6
	Right	41	44.1	14.9
	Catching	27.6	50.9	21.5
	Throwing	34.6	44.4	21
	Sliding	6.1	23.9	69.9

Climbing	19.8	29.6	50.6

	0 trials correct		1 trial correct		2 trials correct	
	M	SD	M	SD	M	SD
Movement verbs (9)	29.42	18.82	32.59	10.15	37.97	19.53
Measurement words (6)	32.83	17.17	31.53	7.75	35.63	18.70
Prepositions (8)	43.21	14.83	40.73	7.24	16.06	9.83
Count words (5)	13.18	8.70	18.24	8.60	68.58	16.08
Operations (2)	49.10	1.70	41.70	5.23	9.20	3.54
Comparatives (4)	40.80	13.85	29.53	2.83	29.65	12.96
Lesson 1: Zoo (1)	26.68	12.44	28.80	9.39	44.52	21.21
Lesson 2: Bakery	33.44	25.46	23.98	13.07	42.56	37.27
Lesson 3: Zoo (2)	32.83	17.17	31.53	7.75	36.13	18.70
Lesson 4: Fruit shop	36.16	20.87	37.06	10.02	26.76	17.88
Lesson 5: Forest	43.45	19.16	32.88	9.01	23.65	18.46
Lesson 6: Playground	28.17	13.42	39.42	10.33	32.42	22.55
Number domain	30.84	17.68	28.35	9.94	40.81	24.77
Space domain	35.91	17.99	36.42	9.59	27.66	18.98

Appendix VI: Van den Berghe et al., submitted

A Toy or a Friend? Children's Increasing Anthropomorphism of a Robot Correlates with Higher Second Language Word Learning

***Abstract*—This paper describes a study that investigated the degree to which children anthropomorphize a robot tutor and whether this anthropomorphism relates to their learning in second language (L2) tutoring sessions. To this end, a robot perception questionnaire was administered prior to and following seven L2 vocabulary tutoring sessions with a humanoid robot. Children tended to anthropomorphize the robot, although they showed large individual differences. As a group, children's responses indicated a slight decrease in anthropomorphism following the L2 tutoring sessions with the robot. However, children's trajectories differed: 20% of the children increased in anthropomorphism, 43% were constant in anthropomorphism, and 37% decreased in anthropomorphism. Further analyses showed that there was a low but significant correlation between change in anthropomorphism and scores on a delayed L2 vocabulary comprehension post-test. We do not know the direction of this relation, but our results show the need to consider children's anthropomorphism when designing robot-assisted tutoring sessions.**

***Index Terms*—anthropomorphism, child-robot interaction, educational robots, second-language learning**

I. Introduction

A. Anthropomorphism

When interacting with a social robot, people have a tendency to attribute human form, characteristics, and/or behaviors to the robot. This phenomenon is called anthropomorphism [1]. People do not only anthropomorphize robots, but also many other nonhuman entities, such as animals, machines, and even natural phenomena [2], and this helps them to gain control over their environment [3], [4]. Anthropomorphism can be a useful mechanism in human-robot interaction [3], [5] as people evaluate robots more positively, collaborate better, and empathize more with robots that are more human-like or display more human-like behavior [6]–[10].

People do not all anthropomorphize robots to the same degree; there are differences between individuals in the tendency to attribute human qualities to nonhuman entities. XXX (blinded for review). *The first two authors had an equal contribution to this paper.

One of the reasons for these individual differences is that people use their own experiences in rationalizing the actions of an object and in reasoning about its mental states [11], and may thus ascribe different mental states to objects depending on their own experiences. Furthermore, people may differ in motivational aspects to anthropomorphize objects, such as loneliness and need for control [12]. People who are dispositionally lonely are found to be more likely to anthropomorphize their pets than people who are not dispositionally lonely, and people who are in need of control are more likely to anthropomorphize unpredictable animals than people who have less need of control [12]. Thus, in human-robot interaction, the degree to which people anthropomorphize robots likely does not only depend on the type of robot used and the behavior that the robot displays, but also on the specific properties and experiences of the person interacting with the robot.

While most robot research on anthropomorphism has focused on adults, it is not a tendency only found in adults. Children of all ages have been found to anthropomorphize robots [13], [14]. Both younger and older children have been found to attribute mental states to robots, even when noticing and discussing machine-like qualities such as the presence of sensors or an adult controlling the robot [13]. Younger children are even found to be more likely than older children to anthropomorphize robots [13], [15]. They are in particular more likely to assign cognitive and affective beliefs to robots, such as the ability to remember people and understand people's feelings [13]. They attribute fewer biological properties or aliveness to robots than older children [16].

B. Change in anthropomorphism

There are indications that children's perception or expectations of robots can change over time. Bernstein and Crowley [17] asked children to judge different entities (including two robots) on liveliness and intelligence. Children who had less knowledge about robots judged the robot more often as alive than children that already had experience with robots. The latter group were more likely to distinguish robots from other entities that they already know (e.g. things that are alive) and judge robots as intelligent; however, not in a human-like manner, but in a unique robot intelligent manner. Westlund and colleagues [18] framed a robot as a social agent or machine by using either inclusive language and second-person pronouns or third-person pronouns and using the word 'robot'. They assessed children's anthropomorphism through a questionnaire both before and after playing a sorting game with the robot. They did not find an effect of framing or having interacted with the robot on children's anthropomorphism. It is not clear whether children's anthropomorphism is not affected by interacting with robots, or whether one interaction was not enough to change their degree of anthropomorphism.

Sciutti and colleagues [19], [20] investigated what children focused on for the design of a robot. They found that the shape of the robot was the primary focus of (young) children before they interacted with the robot, e.g. the robot should contain a head and arms; however, after an interaction with some robots, the shape of the robot became less interesting for the children and the robot's sensory and motor properties became more important, i.e. the robots' ability to feel and move. However, they did not investigate how much children anthropomorphized the robot; they only looked at shape, sensory and motor properties of the robot and how children's expectations changed about an interaction. Even though they did not investigate the way that these properties play a role in child-robot interaction, their

research shows that sensory properties, which can be linked to anthropomorphism, may become important over time.

C. Anthropomorphism and learning

As discussed earlier, anthropomorphizing robots seems advantageous for human-robot interactions [3], [5], but it is not clear how it plays a role in robot-assisted learning. The degree to which learners anthropomorphize robots may play an important role, as learning is first and foremost a social process [21]. The robots' potential for social interactions to establish common ground is one of the advantages social robots have over other forms of technology such as tablet screens. Physical robots indeed have generally been found to be more enjoyable and a preferable social presence than their virtual counterparts [22], [23]. In theory, robots are more natural conversational partners, and may use human-like behaviors such as gestures to support learning. Furthermore, robot gestures have been found to support learning [24]–[26], which suggests that robot-assisted learning interactions benefit from similar social behaviors as human learning interactions, which also benefit from gestures [27], [28]. Findings on the effect of this embodied presence of robots on learning gain as compared to virtual robots, however, are mixed [29], [30].

While robots have clear advantages in theory, it is not clear whether the degree to which learners anthropomorphizes the robot affects how much they learn from it. Children who anthropomorphize the robot more, might interact with the robot similar as they would interact with peers. Literature on peer learning shows possible benefits of peers on learning [31]–[34], and robots may have similar benefits for learning when being treated as a peer. However, it is possible that a robot's benefit depends on the degree to which the learner anthropomorphizes it. This begs the question whether anthropomorphism and learning are related to each other. This has not been investigated yet and is the central research question of our paper.

Research that comes closest is that of Chandra et al. [35], who investigated whether children's perception of a robot, in terms of intelligence, likability, and friendliness, affects their learning in a learning-by-teaching paradigm. Twenty-five seven-to-nine year old children taught a Nao robot to write over the course of four sessions. There were two conditions:

(1) the robot improved its handwriting for half of the children, and (2) did not improve its writing for the other half of the children. Children in the first condition were able to perceive the robot's improvement by the last session, but this did not affect the robot's perceived intelligence, likability, and friendliness. Moreover, their learning was correlated with the likability of the robot. While in the condition in which the robot did not improve, children's learning was only correlated with the friendliness of the robot. These findings need to be interpreted with caution due to the small sample size, but suggest that children's perception of the robot may be related to their learning.

Our study is aimed at expanding previous work in two ways. First, it includes a larger sample. Second, it measures the degree to which children anthropomorphize the robot both before and after having interacted intensively with it, by assessing whether they perceive the robot more as a machine or more as a human. Specifically, we assess the degree to which children assign biological and mental-state properties to the robot, as there seem to be differences in the degree to which children assign these two types of properties to robots [13], [16].

D. This study

This study is part of the project XXX (blinded for review). The current study is part of a large-scale study that evaluates the effectiveness of a social robot in aiding young children's L2 learning. The study included four conditions: (1) robot with iconic gestures (gestures that visualize target words and pointing gestures), (2) robot without iconic gestures (only pointing gestures), (3) tablet-only condition (no robot involved), and (4) control condition (children only danced with the robot). In this paper we only include the experimental robot conditions to investigate children's perception of the robot and the way it relates to their learning. We address the following research questions:

RQ1 To which degree do children anthropomorphize the robot, and does children's perception differ for biological and mental-state properties of the robot? We expect children to differ in the degree they anthropomorphize the robot and we expect large individual differences between chil- dren, in line with research on individual differences in anthropomorphism [12]. Children are likely to attribute cognitive beliefs to the robot [13]. However, we do not know how children will answer the biological questions yet.

RQ2 Do children's perceptions change through multiple L2 tutoring sessions with the robot? There is research suggesting that children's anthropomorphism may not change or decrease upon interacting with robots [17], [18], but we expect that children's perceptions may change over time in different ways in our experiment, due to the multiple interactions children have with a robot. On the one hand, children may perceive the robot as more of a friend after repeated interactions, and change from a machine-like perception to a human-like perception. On the other hand, it is also possible that they have high expectations of the robot's interactive qualities, which the robot cannot meet. In that case, their perception would change from a human-like perception to a machine-like perception.

RQ3 Are children's perceptions of the robot related to their learning of L2 words? We expect that children who anthropomorphize the robot perceive the robot more as a peer learner throughout the tutoring sessions than children who do not anthropomorphize the robot. As learning is a social process [21], children may benefit from perceiving the robot as a peer learner, in line with literature on benefits of peers on learning [31]–[34]. Moreover, there are indications that children's perception of robots is related to their learning [35].

II. Methods

A. Participants

In this study, 119 Dutch children (58 boys) with an average age of 5 years and 8 months (SD = 5 months) participated in L2 (English) tutoring sessions with the robot. They were recruited from nine different kindergartens in the Netherlands. Within kindergartens, they were semi-randomly (taking gender into account) assigned to one of two conditions. There were 62 children (30 boys) in the iconic gesture condition (M age = 5 years and 8 months, SD = 5 months), and there were 57 children (28 boys) in the no iconic gesture condition (M age = 5 years and 8 months, SD = 4 months). During the tutoring sessions, ten children dropped out, of which eight were in the iconic gesture condition. We included their pre-test data but did not obtain post-test data. Furthermore, pre-test data from three children (all in the no iconic gesture condition) who completed the tutoring sessions and the post-tests could not be obtained due to illness. All children's parents signed an informed consent form for their children to participate in this study.

Fig. 1. Experimental setting

B. L2 Tutoring sessions

The aim of the L2 tutoring sessions was to teach each child 34 English words. Each child received seven sessions with the robot and a tablet. The Softbank Robotics NAO robot was used, which was sitting in a 90 degree angle from the child (see Figure 1).

The tablet taught the robot and the child the target words. For each word, the child and the robot had to perform different tasks on the tablet (dragging objects on the screen, repeating target words, or acting out target words). During these tasks, the robot acted as a more knowledgeable peer that was also learning English, but provided feedback on the child's actions when needed. For example when a child was reluctant to drag an object on the tablet, the robot could perform this task for the child. In the iconic gestures condition, the robot used an iconic gesture with every L2 target word, while, in the other condition, it did not. All other gestures were exactly the same across conditions. The complete interaction was autonomous, except for the recognition of children's speech. The interaction was a one-on-one interaction, but the experimenter stayed in the same room to intervene when necessary. During each of the sessions children were introduced to five or six new target words. Prior to the first session, they received a pre-test testing their knowledge of the English target words. After the last session, they performed an immediate post-test within two days of the last session, and a delayed post-test two to five weeks after the immediate post-test (for more details on the study, see [**anonymous**]).

C. Materials and measurements

In this paper we focus on two different measures, each administered twice: 1) a perception questionnaire measuring the degree to which children anthropomorphized the robot, administered once prior to the very first session and one after the seventh and last session 2) a comprehension test mea- suring children's L2 vocabulary learning gain, administered during an immediate post-test and a delayed post-test. Other measurements are beyond the scope of this paper as they assessed other variables predicting children's learning gain but not anthropomorphism (see [**anonymous**] for more results on these measurements) or excluded due to floor effects, i.e. a Dutch-English translation task in which children were able to produce only a few translations.

1) Perception questionnaire: This questionnaire measured to what extent the children perceived the robot as a human or as a machine. The questionnaire was administered by an experimenter and took about ten minutes to complete. It consisted of fourteen questions: seven on the robot's biological properties and the remaining seven on the robot's mental- state

properties (see Table I). The biological and mental- state questions were combined into the children's anthropomorphism score. Each question could be answered with 'yes'/'no'/'I don't know' and had an open-ended query why children answered with this response. The items were based on Jipson and Gelman [16] in which they investigated to what degree children make a distinction between living and non-living items and between biological and mental-state properties of the robot. The children were allotted one point for each 'yes' answer. Two exceptions were the questions on breaking and being made by humans, which were rewarded with one point when answered with 'no'. Thus, the maximum score was fourteen, with a higher score denoting a child's tendency to consider the robot as a human rather than a machine. Cronbach's alpha indicated that the internal consistency of the questionnaire was acceptable, $\alpha = .691$ for the pre-test and $\alpha = .715$ for the post-test.

2) Comprehension test: The comprehension test was a picture-selection task. In this task, children were presented with a prerecorded target word and asked to choose which one out of three pictures matched this word ('Where do you see: [heavy]?'). Each target word was presented three times in a random order. However, only half of the target words were included, as a test including all target words would take too long for these young children. The same test was used for both post-tests. Cronbach's alpha indicated that the internal consistency of the comprehension task was acceptable, $\alpha = .71$ for the pre-test and $\alpha = .75$ for the post-test.

TABLE I
ITEMS OF PERCEPTION QUESTIONNAIRE ASSESSING BIOLOGICAL VERSUS MENTAL-STATE PROPERTIES OF THE ROBOT

Biological properties	Mental-state properties
	Do you think Robin the robot. . .
can see things?	can be sad?
is made by a human?*	can remember
something? can feel it when you tickle?	can think?
has to eat?	understands when you say something?
can feel pain?	can enjoy something?
grows?	can be happy?
can break?*	can recognize you?

*These two questions were allotted one point when answered 'no'.

D. Procedure

Prior to the experiment all children participated in a group introduction with the robot [36]. The robot was introduced to familiarize the participants with the robot, build trust, and explain the similarities and dissimilarities between the robot and humans (e.g. the robot speaks without moving its mouth, but looks at us while speaking the same way humans do). Explanations on the latter were required to make sure that children would know how to interact with the robot. During the introduction, participants danced together with the robot, were allowed to shake the robots hand, and played a brief gesture imitation game. The robot was not explicitly framed as either a robot or a machine, by avoiding pronouns and by being called 'Robin the robot' (i.e., a combination of a human name and the label 'robot'). After the introduction, the first perception questionnaire was administered, together with a few other measurements. In the weeks thereafter, the children had received seven one-on-one tutoring lessons with the robot, after which the perception questionnaire was administered for the second time, together with the immediate comprehension post-test. Finally, the comprehension test was repeated once more in a delayed post-test, between two and five weeks after the lesson series ended.

E. Data Preparation and Analyses

Children's score on the perception questionnaire was the number of points awarded, divided by the number of questions that were administered. Separate scores for the biological and mental-state questions were calculated in a similar fashion, thus, the number of questions answered with 'yes' divided by the number of questions that were administered.

Note that we do not discuss gender in our result section. We ran our analyses with gender as an additional independent variable, but did not find any gender effects.

III. RESULTS

A. Anthropomorphism of the robot

First, we investigated to which degree children anthropomorphize the robot (RQ1). Table II displays the scores that children obtained on the questionnaire, and shows that there are large differences between children in the degree to which they anthropomorphize the robot. As a group, children perceived the robot more as a human than as a machine.

TABLE II
MEAN SCORES ON PERCEPTION QUESTIONNAIRE

	Pre-test (SD)	Post-test (SD)
Mental-state	.73 (.20)	.74 (.22)
Biological	.38 (.20)	.29 (.21)
Total	.56 (.17)	.51 (.18)

A within-subjects ANOVA showed a main effect of property type, F (1,115) = 314.93, $p < .001$, $\eta p^2 = .73$. On average, children ascribed more mental-state properties than biological properties to the robot.

Table III displays per question the percentage of children who have answered the question with 'yes' (or 'no' in case of the questions on breaking and being made by humans). Children agreed more on the mental-states properties than the biological

properties. Children highly agreed that the robot 'can enjoy something','can be happy', and 'can think'. They disagreed more on its sensory abilities, such as 'feeling it when they tickle' and 'feeling pain'.

B. Change in anthropomorphism

Second, we investigated whether children's perception changed due to L2 tutoring sessions with the robot (RQ2). There was a moderate correlation between pre- and post-test scores on the perception questionnaire, $r(106) = .482$, $p < .001$. As Figure 2 clearly shows, children showed large individ- ual differences in their scores on the pre- and post-test. Most children were consistent in the degree to which they anthropo- morphized the robot (46 children), or anthropomorphized the robot less after having interacted with it in the tutoring sessions (39 children). An increase in anthropomorphism also occurred, but was less common (21 children). Crucially, children's change in anthropomorphism seemed independent of the pre- test anthropomorphism score. Figure 2 shows that, from each level of pre-test anthropomorphism, children increased, de- creased, or were consistent in their anthropomorphism of the robot.

We compared children's answers on the perception questionnaire during the post-test to those of the pre-test. A repeated-measures ANOVA showed an effect of time on children's overall perception of the robot, $F(1, 105) = 6.32$, $p = .013$, $\eta p^2 = .06$. Note that this is a small effect. As a group, children anthropomorphized the robot slightly more during the pre-test than the post-test (see Table II).

Next, we investigated whether there were differences for the two types of properties between the pre- and post-test. Children still ascribed more mental-state properties than biological properties to the robot during the post-test, as confirmed by a within-subjects ANOVA, $F(1, 108) = 435.76$, $p < .001$, $\eta p^2 = .80$. Table II suggests that there was a decline for the biological properties, but not for the mental-states properties. A within-subjects ANOVA shows that the main effect of time, $F(2, 104) = 14.39$, $p < .001$, $\eta p^2 = .22$, was indeed significant for the biological properties, $F(1, 105) = 23.68$, $p < .001$, $\eta p^2 = .18$, but not for the mental-states properties, $F(1, 105) = .14$, $p = .710$, $\eta p^2 = .00$. In other words, children mainly changed their perception of the robot with respect to its biological properties.

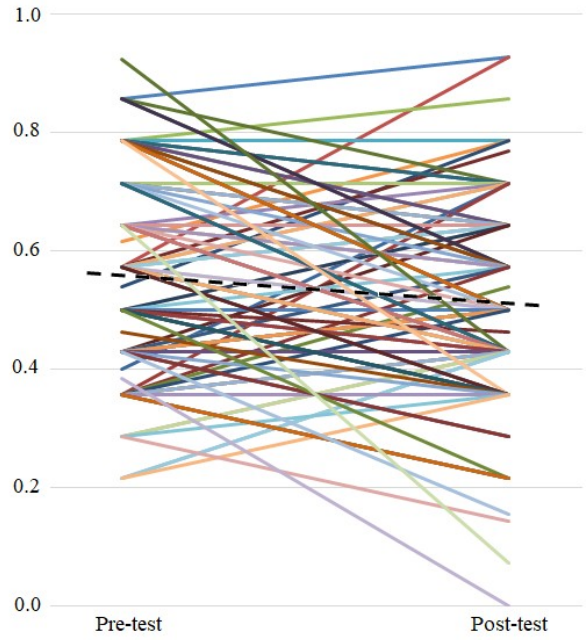


Fig. 2. Individual trajectories from pre- to post-test anthropomorphism. The dashed line represents the group mean.

Table III shows that, for the biological properties, children mainly changed their opinion on the robot's sensory abilities. Fewer children believed during the post-test that the robot could feel it when they tickled or that it could feel pain. With respect to the mental-state properties, more children believed during the post-test that the robot could understand when they said something, and that the robot could recognize them. Fewer children believed that the robot could be sad.

We explored whether children perceived the robot differently in the iconic gesture condition, and the no iconic gesture condition. However, there were no differences in the degree to which children in the two conditions anthropomorphized the robot, $F(1, 104) = .19$, $p = .667$, $\eta p^2 = .00$, and condition did not interact with time, $F(1, 104) = 1.00$, $p = .319$, $\eta p^2 = .01$. Thus, the slight decline in perception was independent of condition, and the use of iconic gestures did not affect the degree to which children anthropomorphized the robot.

C. Anthropomorphism and learning

Last, we investigated relations between children's perception of the robot and their learning during the tutoring sessions (RQ3). We used Pearson's correlations for anthropomorphism scores (pre-test, post-test, and difference scores between pre- and post-test) and comprehension scores (immediate and delayed post-test).

Table IV displays the correlation matrix. We found low but significant correlations between children's anthropomor-

phism and learning. Pre-test anthropomorphism was related to

TABLE III
PERCENTAGES OF 'YES' ANSWERS PER QUESTION

Biological properties	Pre-test	Post-test	Mental-state properties	Pre-test	Post-test
Do you think Robin the robot. . .					
can see things?	78	79	can be sad?	65	51
is made by a human?*	27	15	can remember something?	68	75
can feel it when you tickle?	57	45	can think?	82	78
has to eat?	29	21	understands when you say something?	74	89
can feel pain?	49	39	can enjoy something?	98	96
grows?	20	16	can be happy?	96	98
can break?*	28	25	can recognize you?	64	94

*We reported percentages of 'no' answers for these two questions.

TABLE IV
CORRELATION MATRIX PERCEPTION AND LEARNING

Anthropomorphism	Comprehension	
	Immediate post-test	Delayed post-test
Pre-test	-.213*	-.184
Post-test	-.150	.060
Change	.049	.225*

* Significant at the .05 level.

comprehension scores on the immediate post-test, $r(106) = -.213$, $p = .028$. The relation was negative, suggesting that children who anthropomorphized the robot more prior to starting the lesson series learned less than children who anthropomorphized the robot less. Post-test anthropomorphism was not related to comprehension scores on either post-test, both ps $> .120$.

Children's change in anthropomorphism was related to comprehension scores on the delayed post-test, $r(106) = -.225$, $p = .020$. An increase in the degree to which children anthropomorphized the robot was related to higher performance on the delayed comprehension test. Both the correlation between pre-test anthropomorphism score and immediate post-test comprehension and the correlation between change in anthropomorphism score and delayed post-test comprehension show that children who initially anthropomorphized the robot, performed worse on the comprehension tests than children who anthropomorphized the robot to a lesser degree. Further correlation analyses show that difference scores on the mental-state properties correlated with delayed comprehension scores, $r(106) = .207$, $p = .033$, rather than the biological properties, $r(106) = .160$, $p = .102$. Thus, the degree to which children started to perceive the robot as having mental states correlated with their comprehension scores on the delayed post-test.

IV. DISCUSSION

We investigated whether children perceive a robot more as a human than a machine and whether this is related to children's learning gain in robot-assisted L2 tutoring sessions. We measured the degree to which five- and six-year-old children anthropomorphized the robot both before and after L2 tutoring sessions, and related this to their learning gain on an immediate and delayed comprehension post-test.

Anthropomorphism of the robot

First, we investigated the way children perceived the robot after an introduction session and prior to the tutoring sessions. Children showed large individual differences in their perception of the robot, in line with research on individual differences in the tendency to anthropomorphize objects [11], [12]. Children were more likely to ascribe mental states to the robot than biological properties. Previous work has also found that young children are likely to ascribe cognitive beliefs to robots [13].

In this paper, we did not discuss children's answers to the open-ended questions, which asked them to motivate why they perceived the robot more as a human or machine. However, we saw that there were, similar to their perception scores, large differences in the way children motivated why they perceived the robot in the way that they did. For example, some children thought that the robot would be sad if children did not want to play with it, while other children thought the robot would be sad if it was in pain. Other children thought that the robot could not be sad because it had no feelings while other children thought the robot could not be sad because it could not handle water and, thus, could not cry. The answers to the open-ended questions will be analyzed further to gain a deeper understanding of children's tendency to anthropomorphize robots.

A. Change in anthropomorphism

Second, we investigated whether children's perception of the robot had changed after the L2 tutoring sessions. Overall, children anthropomorphized the robot slightly less after the L2 tutoring sessions than before. We saw an overall decline in the score of biological properties during the post-test and no change for the score of the mental-state questions. Thus, children's anthropomorphism was mainly changed because of the biological properties of the robot, rather than the mental properties. Apparently, the robot had some biological properties that were different from what children thought when first seeing the robot. Fewer children answered 'yes' to questions regarding the sensory abilities of the robot during the post-test as compared to the pre-test: e.g. that the robot 'could feel it when they would tickle', and 'could feel pain'. This is in line with the study of [19] in which they found that the robot's sensory and motor properties became more important for the robot's design after children interacted with a robot. In addition, fewer children thought that the robot 'is made by a human'.

Even though overall scores on mental-state properties did not differ between the pre- and post-test, children changed their beliefs on some mental-state properties of the robot. During the post-test, more children answered 'yes' to questions related to whether the robot can remember something, under- stand them when they say something, and is able to recognize them. We believe that this is due to the setup of the lessons. The robot greeted the children with their name, referred to the previous lessons and tracked the children's faces. It is possible that fewer children believed during the pre-test that the robot could recognize them, simply because they had not played with the

robot in a one-on-one setting yet. Also, fewer children believed during the post-test that the robot 'could be sad', which can also be explained by the design of the lessons: even though the robot expressed its happiness (by changing the colors of its eyes), it did not express negative emotions, like sadness.

Moreover, there were large individual differences between children on their post-test anthropomorphism scores, similar to the pre-test. Most children anthropomorphized the robot to the same or a lesser degree during the post-test as compared to the pre-test. Fewer children increased their anthropomorphism of the robot. It is possible that decreases in anthropomorphism were due to children having high expectations of the robot's interactive (human-like) qualities, which the robot could not meet [37]. The robot was autonomous during the tutoring sessions, and did not engage in personalized conversations with children. It stuck to the script and did not answer their questions. For children with high human-like expectations of the robot, this could have affected the degree to which children believed that the robot interacted with them similar to humans. This also works the other way around, children who perceived the robot as more of a machine prior to the tutoring sessions may have had very low expectations of the robot's interactive (human-like) qualities. Since the robot displayed some human-like behaviors, such as saying the child's name (suggesting that it could recognize the child) or indicating that it liked the sessions, this could have increased the children's human- like beliefs about the robot due to the repeated interactions. Thus, it is possible that the robot's behaviors either encouraged or discouraged anthropomorphism, depending on the user's expectations of the robot prior to the interaction. The way that children perceived the robot and the way that their perception changed, may have been more dependent on their expectations of the robot prior to interacting with it rather than its behaviors or design.

B. Anthropomorphism and learning

Last, we investigated whether children's perception of the robot was related to their learning gain. We found two low but significant correlations. We found that children's anthropomorphism of the robot during the pre-test was negatively related to their comprehension scores on the immediate post-test, and that change in perception was positively related to learning gain on the delayed post-test. Thus, children that increased their anthropomorphism performed better on the delayed post- test than the children who did not change or decreased their anthropomorphism.

Unlike our expectations, only a change in anthropomorphism was related to learning and not the children's pre- and post-test perception. Possibly, this is again linked to children's expectations of the robot. If children had high expectations which the robot could not meet, they may have been disap- pointed with both the robot and the tutoring sessions, which is not beneficial for learning. The robot exceeding expectations may have had a positive effect on children's experiences in the tutoring sessions and their learning.

Note that we could not test the causality of the effect, as there could be a bidirectional relationship between learning and change in perception. It is not clear whether children learn more from the robot because they start perceiving it as a human, or that they start perceiving the robot as a human because they have successful language-learning interactions with it. If children struggled with learning, the robot often had to repeat its requests or had to provide feedback. Both were repetitive behaviors, which likely do not contribute to children's anthropomorphism of the robot.

C. Strengths, limitations, and future research

Our study had several limitations. We did not use a standardized questionnaire for anthropomorphism because of our young target group. Standardized tests such as the Godspeed questionnaire [1] often use Likert scales which are difficult for young children. However, we based our questionnaire on previous work [16] and the questionnaire proved to be reliable. Furthermore, we do not know how the introduction of the robot affected children's perception of the robot. The introduction ensured children would have a common ground when the pre-test perception questionnaire was administered. If the questionnaire had been administered prior to the introduction, it would not have been clear whether the children's answers were based on interactions with similar or different robots, television shows, or imagination. The large variation in scores indicates that children still formed their own opinions about the robot, but we do not know whether children were biased towards anthropomorphizing the robot by the introduction. Furthermore, we do not know whether children had previous experiences with robots, and how individual differences in personality traits such as loneliness or need for control affected the degree to which they anthropomorphized the robot [12].

However, our study also had several strengths. It is one of the first studies to investigate changes in children's anthropomorphism after multiple exposures to robots. Crucially, it also the first to investigate how anthropomorphism relates to children's learning. The different robot property types measured by the questionnaire allowed for a more thorough understanding of the way children perceive robots. Another strength is that we included a delayed post-test to measure children's L2 word learning. Words need time to be consolidated (for a review, see [38]), and learning gains are often assessed better during delayed post-tests than immediate post- tests.

In our study, we could only conduct correlation analyses between children's perception and learning. Future research could explore whether framing the robot as a machine or as

similar to a human affects children's learning. However, anthropomorphism in itself may not be required for successful tutoring sessions, as no positive main effects of anthropomorphism were found in our study. Managing children's expectations of robots, on the other hand, may be crucial. It is an open question whether managing children's expectations prevents them to decrease in anthropomorphism and whether this benefits their learning.

V. Conclusion

The study presented in this paper aimed to explore how children anthropomorphize a humanoid robot, whether their perception had changed after seven tutoring sessions, and whether the change in perception correlated with children's learning gain during these sessions. We found that children generally anthropomorphize the robot, although there were large individual differences in the degree. They ascribed more mental-state properties than biological properties to the robot. Moreover, our results show that children's tendency to anthropomorphize had slightly declined after the tutoring sessions, but their perception trajectories differed: most children anthropomorphized the robot to the same or a lesser degree during the post-test as compared to the pre-test, and fewer children increased their anthropomorphism of the robot. Most importantly, we saw that there was a low but significant correlation between anthropomorphism and learning gain in the delayed post-test: children that started perceiving the robot more as a human learned more from the tutoring sessions. We do not know the direction of this relation, but our results show the need to consider children's anthropomorphism when designing robot-assisted tutoring sessions.

Acknowledgment

This study was carried out within XXX (blinded for review), funded by XXX (blinded for review). We are grateful to XXX (blinded for review) who have helped us conduct this study. We would like to thank the children, their parents, and the schools for their participation. Furthermore, we would like to thank XXX (blinded for review) for their help in collecting data.

References

[1] C. Bartneck, D. Kulic, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, pp. 71–81, 2009.

[2] L. R. Caporael, "Anthropomorphism and mechanomorphism: Two faces of the human machine," *Computers in Human Behavior*, vol. 2, pp. 215–234, 1986.

[3] B. R. Duffy, "Anthropomorphism and the social robot," *Robotics and Autonomous Systems*, vol. 42, pp. 177–190, 2003.

[4] A. Waytz, C. K. Morewedge, N. Epley, G. Monteleone, J.-H. Gao, and J. T. Cacioppo, "Making sense by making sentient: Effectance motivation increases anthropomorphism," *Journal of Personality and Social Psychology*, vol. 99, no. 3, pp. 410–435, 2010.

[5] J. Fink, "Anthropomorphism and human likeness in the design of robots and human-robot interaction," in *International Conference on Social Robotics*. Springer, 2012, pp. 199–208.

[6] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Aug 2005, pp. 708–713.

[7] F. Eyssel, F. Hegel, G. Horstmann, and C. Wagner, "Anthropomorphic inferences from emotional nonverbal cues: A case study." in *RO-MAN*, 2010, pp. 646–651.

[8] F. Hegel, S. Krach, T. Kircher, B. Wrede, and G. Sagerer, "Understanding social robots: A user study on anthropomorphism," pp. 574 – 579, 09 2008.

[9] A. Moon, D. M. Troniak, B. Gleeson, M. K. Pan, M. Zheng, B. A. Blumer, K. MacLean, and E. A. Croft, "Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 2014, pp. 334–341.

[10] L. D. Riek, T.-C. Rabinowitch, B. Chakrabarti, and P. Robinson, "How anthropomorphism affects empathy toward robots," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. ACM, 2009, pp. 245–246.

[11] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: a three- factor theory of anthropomorphism." *Psychological review*, vol. 114, no. 4, p. 864, 2007.

[12] N. Epley, A. Waytz, S. Akalis, and J. T. Cacioppo, "When we need a human: Motivational determinants of anthropomorphism," *Social cognition*, vol. 26, no. 2, pp. 143–155, 2008.

[13] T. N. Beran, A. Ramirez-Serrano, R. Kuzyk, M. Fior, and S. Nugent, "Understanding how children understand robots: Perceived animism in child–robot interaction," *International Journal of Human-Computer Studies*, vol. 69, no. 7-8, pp. 539–550, 2011.

[14] C. Monaco, O. Mich, T. Ceol, and A. Potrich, "Investigating mental representations about robots in preschool children," *ArXiv e-prints*, May 2018.

[15] P. H. Kahn Jr, T. Kanda, H. Ishiguro, N. G. Freier, R. L. Severson, B. T. Gill, J. H. Ruckert, and S. Shen, "robovie, you'll have to go into the closet now: Children's social and moral relationships with a humanoid robot." *Developmental psychology*, vol. 48, no. 2, p. 303, 2012.

[16] J. L. Jipson and S. A. Gelman, "Robots and rodents: Childrens inferences about living and nonliving kinds," *Child development*, vol. 78, no. 6, pp. 1675–1688, 2007.

[17] D. Bernstein and K. Crowley, "Searching for signs of intelligent life: An investigation of young children's beliefs about robot intelligence," *The Journal of the Learning Sciences*, vol. 17, no. 2, pp. 225–247, 2008.

[18] J. M. K. Westlund, M. Martinez, M. Archie, M. Das, and C. Breazeal, "Effects of framing a robot as a social agent or as a machine on children's social behavior," in *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 2016, pp. 688–693.

[19] A. Sciutti, F. Rea, and G. Sandini, "When you are young, (robot's) looks matter. developmental changes in the desired properties of a robot friend," in *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*. IEEE, 2014, pp. 567–573.

[20] M. Obaid, W. Barendregt, P. Alves-Oliveira, A. Paiva, and M. Fjeld, "Designing robotic teaching assistants: Interaction design students and childrens views," in *International Conference on Social Robotics*. Springer, 2015, pp. 502–511.

[21] L. Vygotsky, "Interaction between learning and development," *Readings on the development of children*, vol. 23, no. 3, pp. 34–41, 1978.

[22] C. D. Kidd, "Sociable robots: The role of presence and task in human-robot interaction," Ph.D. dissertation, Massachusetts Institute of Technology, 2003.

[23] A. Pereira, C. Martinho, I. Leite, and A. Paiva, "icat, the chess player: The influence of embodiment in the enjoyment of a game," in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 3*. International Foundation for Autonomous Agents and Multiagent Systems, 2008, pp. 1253–1256.
J. de Wit, T. Schodde, B. Willemsen, K. Bergmann, M. de Haas, S. Kopp,
[24] 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.
[25] M. Macedonia, K. Mu¨ller, and A. D. Friederici, "The impact of iconic gestures on foreign language word learning and its neural substrate," *Human brain mapping*, vol. 32, no. 6, pp. 982–998, 2011.
[26] M. Tellier, "The effect of gestures on second language memorisation by young children," *Gesture*, vol. 8, no. 2, pp. 219–235, 2008.
[27] J. A. De Nooijer, T. Van Gog, F. Paas, and R. A. Zwaan, "Effects of imitating gestures during encoding or during retrieval of novel verbs on children's test performance," *Acta Psychologica*, vol. 144, no. 1, pp. 173–179, 2013.
[28] 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.
[29] J. Kennedy, P. Baxter, and T. Belpaeme, "Children comply with a robot's indirect requests," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 2014, pp. 198–199.
[30] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati, "The physical presence of a robot tutor increases cognitive learning gains," in *Proceed- ings of the Annual Meeting of the Cognitive Science Society*, vol. 34, no. 34, 2012.
[31] A. M. O'Donnell and A. King, *Cognitive perspectives on peer learning*. Routledge: New York and London, 1999.
[32] A. King, "Transactive peer tutoring: Distributing cognition and metacog- nition," *Educational Psychology Review*, vol. 10, pp. 57–74, 1998.
[33] K. Topping, S. Hill, A. McKaig, C. Rogers, N. Rushi, and D. Young, "Paired reciprocal peer tutoring in undergraduate economics," *Innovations in Education and Training International*, vol. 34, pp. 96–113, 1997.
[34] F. Yarrow and K. J. Topping, "Collaborative writing: The effects of metacognitive prompting and structured peer," *British Journal of Educational Psychology*, vol. 71, pp. 261–282, 2001.
[35] S. Chandra, R. Paradeda, H. Yin, P. Dillenbourg, R. Prada, and A. Paiva, "Do children perceive whether a robotic peer is learning or not?" in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2018, pp. 41–49.
[36] P. Vogt, M. de Haas, C. de Jong, P. Baxter, and E. Krahmer, "Child-Robot Interactions for Second Language Tutoring to Preschool Children," *Frontiers in human neuroscience*, vol. 11, no. March, pp. 1–7, 2017.
[37] K. Dautenhahn and I. Werry, "Towards interactive robots in autism ther- apy: Background, motivation and challenges," *Pragmatics & Cognition*, vol. 12, no. 1, pp. 1–35, 2004.
[38] E. L. Axelsson, S. E. Williams, and J. S. Horst, "The effect of sleep on children's word retention and generalization," *Frontiers in Psychology*, vol. 7.