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

D7.6 Integrated report and recommendations

Due Date: **31/12/2018**
Submission Date: **14/01/2019**

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

In this deliverable we first shortly review the main results and challenges encountered during the design and execution phases of the large-scale evaluation study. These results and challenges have extensively been reported in previous deliverables. Therefore, here we only shortly review them and refer to the corresponding deliverable. Following this, we report about two additional analyses done with the data of the large-scale evaluation study.

First we report about an analysis where we examined if skills that are known to be important for second language learning (i.e., vocabulary in first language, phonological memory, and selective attention) moderate the results previously reported. Similar results were obtained as in previously reported analyses: children in the experimental conditions outperformed those in the control condition, but there were no differences between experimental conditions. Results show that children with higher levels of Dutch vocabulary and selective attention learned more L2 words when they were in the experimental conditions (i.e., the difference between the experimental and control condition is larger for children with larger Dutch vocabularies and/or better selective attention). Moreover, children with better phonological memory and larger Dutch vocabularies seem to learn more L2 words in the condition of the robot that does not perform iconic gestures than in the condition of the robot that does perform these iconic gestures. Children with better selective attention show the opposite trend (i.e., better performance in the iconic gesture condition than in the no-iconic gesture condition), at least when word knowledge was tested with translation tasks rather than a comprehension task. Some analyses also suggest that children with better Dutch vocabulary and/or phonological memory learn more words in the robot-assisted conditions than in the tablet-only condition and that for children with better phonological memory this effect is opposite. However, these results were only found for the translation tasks and not for the comprehension task.

Initial descriptive results of data about engagement (not part of the scope of the L2TOR study) suggest that in the tablet-only condition children show higher task engagement than in the two robot conditions. Robot engagement seems to be slightly higher in the robot with iconic gestures condition. Moreover, both task and robot engagement seem to decline over time. The decline in task engagement is also evident in the tablet-only condition (though we do not have sufficient data yet to see the trend over the entire lesson series). This suggests that the novelty effect might not be unique to the robot conditions, but a more general effect for learning with technological aids.

In the final section of this deliverable we provide a comprehensive discussion of the results of the large-scale evaluation study. We discuss the lessons learned, draw conclusions and provide directions for future research as well as recommendations for areas where social-robots might effectively be used in an educational context.

Principal Contributors

UU: Rianne van den Berghe, Esmee Kramer, Ora Oudgenoeg-Paz, Paul Leseman
TIU: Mirjam de Haas, Jan de Wit, Emiel Krahmer, Paul Vogt, Bram Willemsen

Revision History

Version 1.0 (RB 14-01-2018)
This is the first version.

1 Introduction

This deliverable is intended to serve as an integrated report on the results of the large-scale evaluation study. Previous deliverables have reported in detail about the experiences during the design phase, the technical issues encountered, and some first results of the study. In this deliverable we first provide a short overview of these reports, with references to the corresponding deliverables. Following this, we report about two follow-up analyses done with the data of the large-scale evaluation study that were not yet included in previous deliverables. Specifically, in section 2, we report on analyses of individual differences in language and attentional skills as related to the learning outcomes in the different experimental domains. We tested whether children's phonological memory, vocabulary in L,1 and selective attention (which are all skills that are known to support language learning) moderate the effect of the experimental conditions on learning gains.

In section 3, we report on the initial descriptive analyses of engagement during the lessons series. This coding of engagement is an additional analysis performed by the partners in UU and TIU, which falls outside the scope of the L2TOR project and is still ongoing. However, as the results are interesting to report here, we discuss the initial results. We coded a small subset of the films recorded during the lessons series and scored children's engagement with both the task and the robot. In this analysis, we compare both forms of engagement between the different experimental groups and describe changes in engagement across the lessons series. Finally, in section 4, we provide an integrative discussion of all the findings of the large-scale evaluation study, offer possible explanations, and make recommendations for scientists and professionals in the field of education.

After the review meeting, in line with the recommendations of the reviewers, the design of the large-scale evaluation study as specified in the proposal was adjusted. The scale of the study was reduced to include only one language combination, namely Dutch children learning English. Additionally, it was decided to include two types of teaching interventions with the robot. Specifically, a robot using only deictic gestures and a robot using deictic and iconic gestures were included. There were two comparison groups: a group learning only with a tablet and a control group doing dances with the robot and not learning any English.

The content for the lessons was developed based on existing curricula and extensive observations of human teaching of L2. These developments have been detailed in deliverables 1.1 and 1.2 and 1.3. In line with the revised objectives, where we decided to focus on iconic gestures in one of the experimental groups because of the benefits found in previous work, in D1.3 we focussed specifically on the selection of suitable gestures to be used in the study.

The content developed was implemented in the L2TOR system. The partners who did the implementation (PLYM, UNBI, TIU, ALD) were in close contact with the partners who developed the content (UU, KOC). Based on issues encountered during the implementation, the content was adjusted. This process continued till the start of the large-scale evaluation study to ensure that the lessons are implemented in the best manner possible technologically, but also that educationally sound content was provided that will enable children to learn.

The main limitation that required the adjustment of the lessons was clear already early in the project. Specifically, we concluded that the current technology for Automatic Speech Recognition (ASR) and object recognition is not yet sufficient for implementation within the L2TOR System (see D4.1). Therefore, it was decided to work with a tablet that functions as a mediator. The child performs actions on the tablet and according to these actions the robot can respond and provide feedback, if necessary. Additionally, due to the voice activation technology still being unreliable, we decided to have a human experimenter using a laptop to indicate whether children repeated the words in L2 when asked to do so during the lessons. Note however, that apart from this, the L2TOR system functioned autonomously. See D4.3 for the final specifications of the input modules of the L2TOR system.

WP2 worked on translating the lessons into storyboards that specified the interactions for both content domains (i.e., the number and space domains). D2.1 and D2.2 show our initial versions for interaction specifications for both domains. D2.3 shows the final versions that were adjusted based on the design decisions taken due to technical limitations and following the revised objectives. In order to be able to compare the robot-assisted and tablet-only conditions, the interactions were adjusted such that text spoken by the robot could also logically be used in the tablet only condition, while not losing the possible advantages of the robot.

In the deliverables of WP6, we extensively evaluated the challenges we encountered with the setup of the interaction in the large-scale evaluation study. See D6.3 for an evaluation of the usability of the tablet interface and D6.4 for a discussion of the challenges regarding multimodality, speech synthesis and the use of tablet-mediated interaction.

In D7.1 we described the design of the large-scale evaluation study in detail. The design is different than the design originally stated in the proposal, following the revised objectives. The aim of this large-scale evaluation study was to conduct a randomised controlled trial with sufficient power to test the effectivity of a robot peer-tutor for teaching Dutch speaking 5-6 years old children English words in the domains of early mathematics and spatial language. This required employing a strict protocol including pre- and post-tests and defining clear rules for the role of the experimenter, timing of the lessons and tests, and so forth. This study is the first study in this field to employ a sample with sufficient power, and a strictly controlled design as is common in the educational studies. As such, the protocol, included in D7.1, can be used to guide future studies in this field. This is highly important, as it is essential that studies on employing social robots in educational contexts adhere to common scientific norms in the field of education. Moreover, the hypotheses of the study were pre-registered, thus ensuring transparency and testing of the hypotheses as they were formulated prior to data collection and analyses.

In D7.2 we presented the analyses of the pre-registered hypotheses showing that while children in all of the experimental groups did learn L2 vocabulary (as compared to the control group), no difference was found between children in the three experimental conditions. Thus, the use of iconic gestures by the robot did not appear to support L2 learning more than the use of only deictic gestures, and children did not learn more when learning with the robot and a tablet than when learning only with a tablet. In D7.2 we discussed the meaning of these results and suggested some possible explanations.

In this deliverable we report on two additional analyses aiming to further explore the data of the large-scale study and look for possible moderators playing a role. In D7.5 we specifically discussed the comparison between the tablet-only and robot-assisted conditions and discuss the reasons why we do not see any difference between these conditions. We mainly focus on the idea that, in the current setup with the prominent role played by the tablet, the added value of the robot was minimal. It might even be that (especially in the robot with iconic gestures condition) the robot distracted the children. We further address this issue also in section 3 of this deliverable where we report about our coding of children's engagement with the task and with the robot. Moreover, in D6.3 we elaborate more about the comparison of the two robot conditions (with and without iconic gestures). We discuss the reasons why, contrary to our hypothesis, no advantage of the iconic gestures was found. We focus on the setup of the study (children sitting next to the robot) which might have impaired children's view of the gestures. However, this design might have encouraged more repetition of the gestures. Additionally, we discuss the design of the gestures, which was challenging for some target words, individual differences in the effect of the gestures, and the fact that the use of the tablet might have prevented children from benefiting from the iconic gestures. This issue is also further addressed in section 3 of this deliverable, where our analyses of engagement data is presented.

As mentioned above, in this deliverable we first report about some additional results regarding individual differences and engagement with the task and with the robot and end with a through discussion of the results of the large-scale evaluation study and recommendations for future directions that should be pursued in order to enable social robots to become effective language tutors.

2 Individual differences in learning gains

2.1 Introduction

In this section, we report on analyses of individual differences in children's language and attentional skills as related to their learning outcomes in the large-scale evaluation study. In the analyses reported in this section, we investigate whether children benefit from the presence of a robot and from the robot's gestures when learning L2 vocabulary, and whether individual differences in language learning-related skills moderate any beneficial effects of the robot. The design of the study and the analyses of the pre-registered hypotheses regarding differences between experimental conditions and possible explanations for our results were extensively discussed in D7.1 and D7.2.

Thus far, individual differences in robot-assisted language learning have rarely been studied. It is possible that robots are useful L2-education tools for certain children only, for example, children who are either good or poor language learners. We tested whether children's phonological memory, vocabulary in L1 and selective attention moderate the effect of the experimental conditions on learning gains. The data on individual differences will be reported in a paper by van den Berghe et al., which is currently in preparation.

L1 vocabulary knowledge reflects both child-factors such as phonological memory and selective attention (Baddeley, Gathercole, & Papagno, 1998; Gathercole & Baddeley, 1990) and environmental factors such as socio-economic status and quality and quantity of parental input (Hoff, 2003; Rowe, 2013). L1 vocabulary knowledge can benefit L2 word learning in two ways: (1) indirectly, through factors underlying L1 knowledge, such as phonological memory or selective attention or (2) directly through the learner making use of their lexical network in their L1 to learn L2 (e.g., certain concepts or words that are similar in the L1 and L2 are learned more easily).

Phonological memory concerns the ability to temporarily construct a phonological representation of unfamiliar sound sequences in working memory (Gathercole & Baddeley, 1990; for a review on the relationship between phonolgocial memory and word learning, see Gaterhcole, 2006). Phonological memory has been found to contribute both to L1 and L2 vocabulary learning (Baddeley et al., 1998; Gathercole, 2006; Gathercole & Baddeley, 1990; Masoura & Gathercole, 2005; Service, 1992; Verhagen & Leseman, 2016).

Selective attention concerns the ability to focus on a particular object while tuning out unimportant details. It helps the learner to process language, in particular speech segmentation (Stevens & Bavelier, 2012). Thus, it helps the learner to identify words in speech streams. Moreover, unlike L1 acquisition, L2 learning is not necessarily an unconscious process. Therefore, consciously paying attention may be necessary to learn the L2 (see for example the Noticing Hypothesis; Schmidt, 1990).

Thus, L1 vocabulary, phonological memory and selective attention are all suggested in the literature as important factors supporting L2 learning. However, as noted, empirical evidence regarding differential effects of robots is lacking. We decided to focus on these three possible moderators, given the evidence showing their importance for L2 learning. As the domain of individual differences in the effects of robots on learning is still relatively unexplored, we did not formulated any hypotheses, but rather conducted an exploratory analysis of individual differences. Specifically, we studied whether the three

factors (L1 vocabulary, phonological memory and selective attention) moderate the differences between the four conditions of the large scale evaluation study (i.e., control, tablet only, robot without iconic gestures, and robot with iconic gestures).

2.2 Measures

The design and procedure of the large-scale evaluation study were described in detail in D7.1. In this analyses, we used the outcome variables described in the analyses reported in D7.2, namely the two translation tasks (from Dutch to English and from English to Dutch) and the comprehension task. See D7.1 and D7.2 for a detailed description of these tasks. We used the data from both post-test (i.e., the post-test done 1 to 2 days after the last lesson and the delayed post-test done 2 to 4 weeks after the last lesson). Additionally, we used three measures for the three moderators. Below we provide more details about these measures.

Dutch Vocabulary. We used the Dutch version of the Peabody Picture Vocabulary Test (PPVT) to measure children's Dutch receptive vocabulary knowledge (Dunn, Dunn, & Schlichting, 2005). This task is a picture-selection task in which children are presented with four pictures and have to select the picture corresponding to a word said by the experimenter. The task contains a total of seventeen sets, of which each set consists of twelve items. The test is adaptive, such that the starting set is chosen depending on the age of the child, and testing is stopped when the child makes nine or more errors within one set. Besides raw scores, this test has norm scores with an average of 100 and SD of 15. We used the norm scores in our analyses.

Phonological memory. The Quasi-Universal Nonword Repetition Task (Q-U NWRT) was used to measure phonological memory (Boerma et al., 2015; Chiat, 2015). The Q-U NWRT is a computerized task appropriate for young children, consisting of twelve items. Children hear a previously recorded, non-existing word via a laptop computer, and are asked to repeat it. Children receive two practice items (two one-syllable nonwords) before starting. Children's responses were scored online by the experimenter and received one point for each word that they repeat correctly, yielding a maximum score of twelve.

Selective attention. A computerized visual search task was used to measure selective attention (Mulder, Hoofs, Verhagen, van der Veen, & Leseman, 2014). In this task, children were shown a display of animals on a laptop screen and were asked to find as many elephants as possible among distractor animals. Children were given three practice items and four test items that increased in difficulty. In the first two items, 48 elephants, bears, and donkeys (similar in color and size) appeared on a six by eight grid. In the third item, 72 elephants, bears, and donkeys (similar in size to the first two test items) appeared on a nine by eight grid. In the last item, 204 elephants, bears, and donkeys (smaller in size than in the other three test items) appeared on a 12 by 17 grid. There were eight elephants in total in each test item. Each test item lasted 40 seconds. Throughout the test, children were encouraged to search as quickly as possible. Elephants that were found were crossed off with a line by the experimenter. The number of targets located correctly per round were calculated and averaged across items, resulting in a maximum score of eight.

2.3 Analyses and results

To investigate differences in learning gains between the three conditions, we ran linear mixed-effect logistic regression models in the statistical package R (R Core team, 2017) and the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). Dependent variables were children's binary (correct/incorrect) scores on the translation tasks and the comprehension task. The analyses were run separately for the translation tasks and the comprehension task, as they measure different types of vocabulary knowledge. We included 'subjects', 'target words', and 'test item number' as random factors, and random slopes for subjects (subject*post-test). We employed the method of model comparisons, in which the most parsimonious model that best fit the data was identified. Three such models were constructed:

1. A model with the English-Dutch translation task and all three assessment points (the pre-test, immediate post-test, and delayed post-test)
2. A model with both translation tasks and the two assessment points in which they were both administered (the immediate and delayed post-test)
3. A model with the comprehension task and the two assessment points in which it was administered (the immediate and delayed post-test)

Note that the English-Dutch translation task was the only task administered also during the pre-test. Using these three models allowed us to investigate whether children improved from pre-test to post-test, between the immediate and delayed post-test, and differently depending on the different type of test (translation or comprehension). Moreover, it enabled us to test if the effects of time and condition were moderated by (one or more out of) the three moderators.

In these models, the factors 'condition' (control, tablet-only, robot without iconic gestures, and robot with iconic gestures) and 'post-test' (immediate and delayed) were included as fixed effect factors, with an interaction between them. For the translation task, 'language' (from English to Dutch or from Dutch to English) was included as an additional factor. We added each of the language and attentional skills (i.e., 'vocabulary knowledge', 'phonological memory', or 'selective attention') as a fixed effect factor, to investigate whether there were any main or interaction effect of in interaction with condition and post-test. Below, we only report the main outcomes of these analyses. The detailed results with statistics can be found in Appendix B.

First, as discussed before in D7.2, several main effects were found. In all models, we found a main effect of condition, with children in the experimental group outperforming children in the control condition. There were no differences between the three experimental conditions. Furthermore, a main effect of post-test was found, with children obtaining higher scores on the delayed post-test than in the immediate post-test. Note, however, that this effect was only found for the translation tasks and not for the comprehension task. Last, a main effect of language was found, with children obtaining higher scores on the English-to-Dutch translation task than on the Dutch-to-English translation task. A main effect of post-test was found for the translation tasks but not for the comprehension task. It is not clear why no effect was found for the comprehension task in contrast to our previous analyses, but it suggests that the effect of time is not stable for this task.

Most analyses showed an interaction between skill and condition: there were positive effects for L1 vocabulary, phonological memory, and selective attention, but only for children in the experimental conditions, and not for those in the control condition. This interaction was to be expected, as these skills are known to benefit L2 learning and, in our study, only children in the experimental conditions received an L2 vocabulary training in which they could benefit from these skills. Note that these effects were only found for the translation tasks and not for the comprehension task.

Some analyses suggested differences between the robot-assisted and tablet-only conditions for two of the three moderators (i.e., L1 vocabulary and phonological memory). Children with larger L1 vocabularies learned more words in the robot-assisted conditions than in the tablet-only condition, while this effect was opposite for phonological memory: children with better phonological memory learned more in the tablet-only condition than in the robot-assisted conditions.

Other analyses suggested differences between the two robot-assisted conditions. Children with larger L1 vocabularies and/or better phonological memory learned more in the condition in which the robot did not use iconic gestures than in the condition in which it did. Selective attention showed an opposite pattern: children with better selective attention learned more words in the condition in which the robot used iconic gestures than in the condition in which it did not. Thus, differential effects were found for the three language and attentional skills.

2.4 Discussion

The results from these analyses were generally similar to those reported in D7.2: (a) children in the three experimental conditions outperformed children in the control condition; (b) there were no significant differences between the three experimental conditions; (c) children obtained higher scores on the English-to-Dutch translation task than on the Dutch-to-English translation task; and (d) children obtained higher scores on the delayed post-test than in the immediate post-test. Note that in the current analyses, we only found the effect of post-test for the translation tasks and not for the comprehension task. However, given that we used different analyses for this deliverable than in D7.2 (MANOVAs vs linear mixed-effects modelling) the fact that we generally found similar results show that our results are quite robust.

Various effects were found for the language and attentional skills. Note however that these effects were not always consistently found across models and skills, so they must be interpreted with caution. L1 vocabulary, phonological memory, and selective attention benefited children's learning in the three experimental conditions, but not in the control condition. As discussed, this interaction was to be expected, as these skills benefit L2 learning and, in our study, only children in the experimental conditions received an L2 vocabulary training in which they could benefit from these skills. The lack of benefits in the control condition shows that the three skills did not benefit general tests performance on the tasks during the post-tests, but rather, that the skills benefited children's ability to learn from the vocabulary training and therefore their performance on the post-tests.

Differential effects were found for the language and attentional skills with regards to the experimental conditions. Below we offer some interpretations. Note however, that as these effects are not always consistently shown and as this study is the first to test such

effects, these interpretations are somewhat speculative and should therefore be treated with caution. Replication of such findings is required before any firm conclusions can be drawn. First, when comparing the tablet-only and robot-assisted conditions it seems that children with better L1 vocabulary benefit more from the robot-assisted conditions while children with better phonological memory benefit more from the tablet-only condition. It could be that the presence of the robot supports the mapping of L2 concepts to known concepts in L1, whereas without the robot children can more easily concentrate on the auditory stimuli, enabling them to learn the words uttered by the tablet.

Second, when comparing the two robot-assisted conditions, L1 vocabulary and phonological memory benefited learning in the robot without iconic gestures condition in particular. Children with larger L1 vocabularies and/or better phonological memory benefited from the robot's presence, although they did not need its iconic gestures to learn the target words. In contrast, selective attention particularly benefited learning in the iconic gestures condition. The iconic gestures condition is highly demanding in terms of attention (i.e., the learner has to focus on the tablet, robot itself, and its gestures), and good attentional skills may be required to benefit from the robot's gestures. At the same time, the iconic gestures might distract children and hamper their ability to use the L1 vocabulary knowledge and phonological memory to facilitate L2 learning. This would explain why children with better L1 vocabulary and phonological memory performed better with the robot without iconic gestures.

Taken together, the results suggest that the study of individual differences and moderators of the effects is highly relevant. It is very likely that the effects of the robot are different for different children. As noted above, these results should definitely be replicated before any conclusions can be drawn. However, these results call for future studies to include such individual factors and look for differential effects. The study of such individual differences is standard practice in educational sciences and developmental psychology and should therefore be included in studies looking at the implementation of robots for educational practices.

3 Task and robot engagement during the lesson series

3.1 Introduction

In this section, we report on a preliminary descriptive analysis of children's engagement during the lessons. Two forms of engagement were analyzed: engagement with the task, and engagement with the robot. Although much of the instructions given in the task had to be performed on the tablet (e.g., to drag animals into cages), it is important to note that task engagement is not equivalent to tablet engagement. In lessons 5 and 6, for instance, children were instructed to act out verbs such as running and jumping, and this is also part of the task. Moreover, engagement with the robot (in the robot-assisted conditions) when it provided educational content is also seen as task engagement. The current analyses were done on a preliminary subset of all the films that were coded up till the moment of analyzing. This concerns about 10% of the films recorded during the lesson series. Coding of engagement is highly time consuming and requires a large investment in terms of developing a reliable coding scheme, training coders, and doing the actual coding with appropriate double coding procedures in place. The coding and analyzing of all the films falls outside the scope of L2TOR, and will be done by the partners in TIU and UU after the completion of the L2TOR project, using their own resources. However, as the results are potentially interesting to report here, we chose to report the initial results, based on the coding done up to this moment in this deliverable. In future publications we will be able to provide more concrete answers to the research questions presented here.

Engagement is an important factor in robot-assisted language learning research, as engagement is often positively related to learning outcomes (e.g., Konishi, Kanero, Freeman, Golinkoff, & Hirsch-Pasek, 2014; Zaga, Lohse, Truong, & Evers, 2015). That is, the more engaged a child is, the more prone the child is to learn something. Therefore, the lessons were designed in a such a manner that the engagement of children would be encouraged. For example, for the lessons an overall theme was chosen that was familiar and appealing to children. In addition, the robot was introduced as a peer instead of as a teacher, based on previous research with older children demonstrating that this can lead to higher engagement (Zaga, Lohse, Truong, & Evers, 2015).

Kanero et al. (2018) and van den Berghe et al. (2018; see also D7.3) stated in their reviews that children generally enjoy learning with a robot. However, since most studies only include short-term interventions with the robot, the high engagement of children might also be because of a novelty effect (Kanda, Hirano, Eaton, & Ishiguro, 2004; Leite, Martinho, & Paiva, 2013). The current study is able to examine the engagement with both the task and the robot for an extended time during the lesson series, and thus examine the novelty effect in terms of engagement.

In addition, it is interesting to examine engagement patterns in the different conditions. De Wit et al. (2018) found that children interacting with a robot that performed iconic gestures seemed to be more engaged than children interacting with a robot that did not perform such gestures. Task engagement of children in the tablet-only condition is also taken into account in the current study, to be able to examine the (change in) engagement of children who did not work with the robot, and compare this to the robot-assisted groups. This might give even more insight into the novelty effect, as this indicates whether a potential decline in engagement is something specifically related to the robot or a more general effect of (lessons with) technological devices.

With the current analysis we aim to look at the following three questions:

1. Is there a difference in task- and robot engagement between the conditions?
2. Is there a change in in task- and robot engagement over the time of the lesson series? (novelty effect)
3. Is this change over time different between conditions?

Although only a subpart of the video's was coded, the data give a first impression of the engagement data. The way in which engagement was coded is described in more detail below. Subsequently, the results of the preliminary descriptive analysis are reported and discussed.

3.2 Method

Participants

Engagement data was available for *N* = 112 children at the moment of analysing. As described before, this is only a small subset of the total sample, so results should be interpreted with caution. This sample consisted of 57 boys and 53 girls, with a mean age of 5 years and 7 months (*SD* = 4 months). Children from all schools in the study were included, but there was a substantial range in number of children from each school (i.e., from 1 to 33 children). The division over the experimental conditions was as follows: 28 children from the tablet-only condition, 44 children from the condition without iconic gestures, and 38 children from condition with iconic gestures. Taken together, the current sample seems to be a quite random and a representative subsample of the total sample as reported on in D7.2.

Coding method

Film recordings were made of every lesson of all children that participated in the large-scale evaluation study. From every lesson, 3 fragments of 2 minutes each were selected. This was done in a way that one fragment was always roughly from the beginning of the lesson, one in the middle of the lesson, and one from the end of the lesson.

The coding schemes for both task- and robot engagement are attached in appendix A (in Dutch). A group of 8 coders worked together to code the engagement data. They followed a joint training prior to coding and had biweekly skype meetings to discuss cases where they encountered difficulties. The ZiKo evaluation scheme (Laevers, 2005) formed the basis for developing the current coding scheme. This coding scheme involves 5 levels of engagement, where 1 is extremely low and 5 is extremely high. For the L2TOR project, the scheme was extended to include particular situations in the lessons and differences between the two types of engagement (task or robot). Table 1 displays a general description of each level of engagement.

We made a distinction between the two types of engagement because some children were very much engaged with the robot but completely ignored the task or the other way around. Also, this would ensure that task engagement was coded similarly for the tablet-only and robot-assisted conditions, and that differences in engagement scores across conditions were not due to differences inherent to the design of the conditions. Task engagement and robot engagement are not mutually exclusive. That is, children who score high on robot engagement can also score high on task engagement.

For robot engagement, the score depended on how engaged the child was with the robot. If a child ignored the robot, this was coded as low engagement (i.e., a 1). If the child

was looking and responding to the robot, but was also distracted easily, this was scored mediocre (i.e., a 3). Highly engaged children scored high (i.e., a 5), which indicated that the child was interacting with the robot, for instance, reacting to the robot even when the robot was not specifically asking to repeat him. In the iconic gesture condition, an indication of high engagement was when the child mimicked the robot's gestures.

For task engagement, children were rated on how engaged they were with the task. So, if they were not doing the task at all, they got 1 point. If the child was doing the task, but also got distracted easily (looking away, not responding directly, etc.), they got 3 points. If the child was continuously working on the task and the coders saw that the child enjoyed doing the task and was fully concentrated, the child was rated as 'highly engaged'.

A coding tool was created by one of the researchers (see Figure 1), which showed the front-view and side-view of the video (if available), the coding scheme, a place to type comments and the rating scale. The tool was made in such a way that the coder had to view the whole video before he or she could rate it. The raters could re-watch parts of the video after watching the complete video once.

Table 1
General description of each engagement level.

Level	Description
1. Extremely low	The child shows practically no task-related activity/interaction at all.
2. Low	The child shows some activity/interaction, but the activity/interaction is regularly interrupted.
3. Mediocre	There is continuous activity/interaction, but the child is not really concentrated.
4. High	There is continuous activity/interaction, and the child is generally concentrated, but can be distracted.
5. Extremely high	There is uninterrupted activity/interaction, and the child is strongly committed to the activity/interaction.

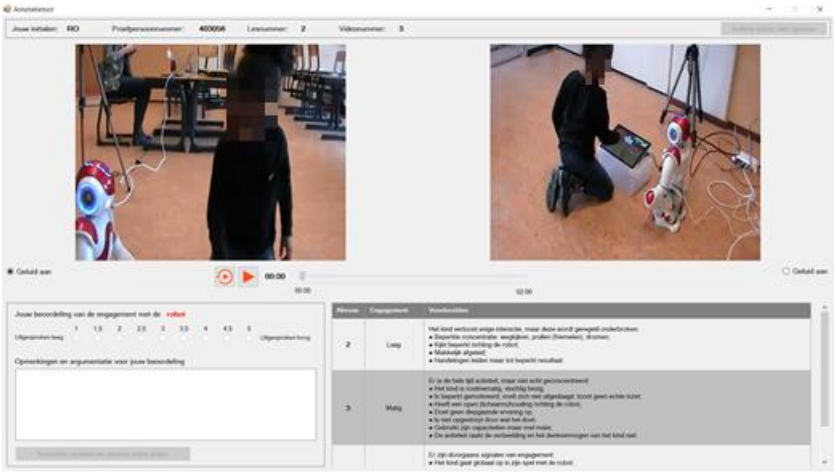

Figure 1. Coding tool

3.3 Analyses and results

The engagement scores of all available fragments from one lesson (max. 3, min. 1) were aggregated at the child level. This was done for both task- and robot engagement. In this way, all children obtained for each lesson one engagement score for task engagement and one engagement score for robot engagement.

Table 2 displays for each condition the mean engagement scores per lesson, along with the number of scores on which these means are based. These results are also graphically presented in Figure 2. As can be seen, more task engagement scores were available compared to robot engagement scores. No mean scores were computed if there were less than 10 scores available, as this might yield a substantial bias due to the low number of observations. Because of the small sample size, only descriptive analyses were performed. Analyzing the current engagement data in relation to the learning gains was also not considered feasible because of the relatively small amount of data. Especially the data that was available of the tablet-only condition was very limited up until the moment of data analyzing.

Table 2
Mean engagement scores per lesson for each condition, together with the accompanying standard deviations and number of scores

Lesson	**Tablet-only**		**Without iconic gestures**		**With iconic gestures**	
			Task engagement			
	M (*SD*)	*N*	M (SD)	*n*	*M* (*SD*)	*n*
1	4.53 (0.30)	21	4.11 (0.50)	37	4.10 (0.58)	33
2	4.10 (0.67)	21	3.85 (0.51)	37	3.65 (0.63)	33
3			3.74 (0.70)	31	3.61 (0.89)	20
4			3.84 (0.64)	30	3.23 (0.69)	17
5			3.47 (0.66)	28	3.01 (1.03)	17
6			3.28 (0.61)	26	2.76 (0.77)	23
7			3.81 (0.55)	16	3.44 (0.70)	17
			Robot engagement			
			M (*SD*)	*n*	*M* (*SD*)	*n*
1	NA	NA	3.41 (0.70)	36	3.64 (0.87)	32
2	NA	NA	3.07 (0.49)	38	3.38 (0.72)	33
3	NA	NA	3.06 (0.70)	30	3.55 (1.00)	12
4	NA	NA	2.88 (0.79)	30	2.89 (0.58)	18
5	NA	NA	2.75 (0.68)	14		
6	NA	NA	2.76 (0.75)	27	2.69 (0.73)	24
7	NA	NA				

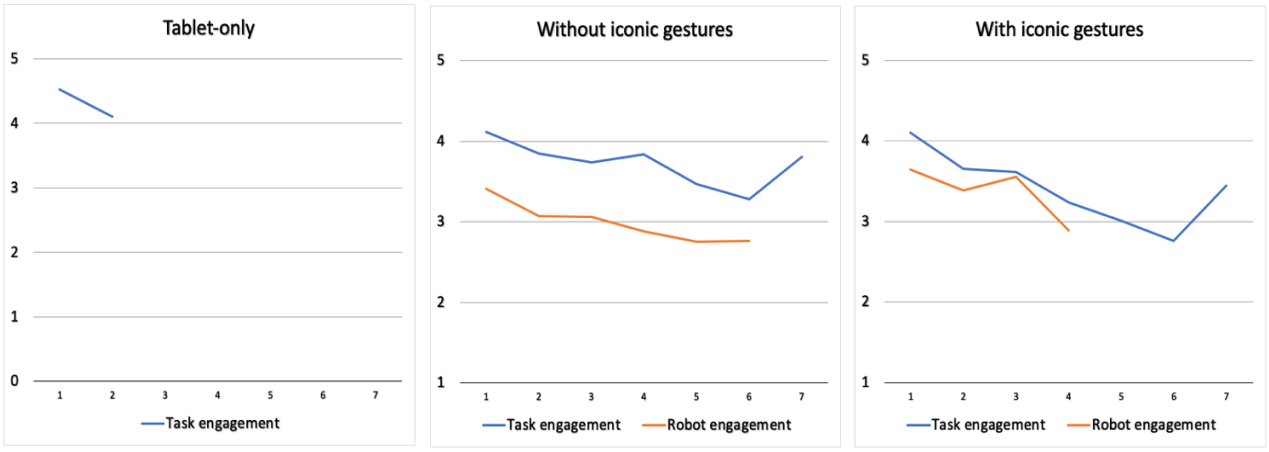

Figure 2. Mean task and robot engagement scores per lesson for the experimental conditions (if more than 10 scores were available)

As these results are based on only a subset of the data, no firm conclusions can be drawn, nor clear answers can be given to the research questions. What we can do, however, is describe the patterns that we see in these preliminary results, which might be an indication of what the results with the complete data look like.

First, it can be noted that the task engagement seems higher in the tablet-only condition for the first two lessons than in the two robot conditions. If this pattern would remain for all lessons, this may suggest that the presence of the robot distracts children from engaging with the task. However, this may also particularly be the case for the first lessons, as children in the robot conditions still have to get used to the robot. Moreover, task engagement is fractionally higher in the robot condition without iconic gestures compared to the condition with iconic gestures. These differences are very small and therefore might not be meaningful. However, they are stable across all lessons coded. If these differences would still be visible and statistically significant in the final analyses, this would support our idea that the iconic gestures distracted the children even more than just the robot without these gestures.

In addition, robot engagement is for all lessons lower than the task engagement, although the differences are small in the iconic gestures condition. Robot engagement thus seems higher in the iconic gestures condition compared to the condition without the iconic gestures. This indicates that the use of iconic gestures might lead to higher robot engagement, which is in line with De Wit et al. (2018). The level of robot engagement in the condition without iconic gestures can be considered relatively low, as a score of 3 indicates that the child is not really concentrated on the interaction, and most of the lessons have an average score around or lower than 3.

Concerning the change in engagement over time, a slightly decreasing pattern is visible for both task and robot engagement. This might point to a novelty effect of the robot, indicating that as time passes by children get accustomed to the robot, and the novelty and accompanied engagement wear off (in line with Kanda, Hirano, Eaton, & Ishiguro, 2004; Leite, Martinho, & Paiva 2013). Only in lesson 7, the recap lesson, task engagement increases again in both robot conditions. This might be because children knew

that it was the final lesson, and were therefore possibly more engaged compared to other intermediate lessons. It should also be noted that lesson 7 was a different lesson than the other lessons, as children did not have to perform tasks like repeating the robot, and they revisited all places and content from the previous lessons. This might have also contributed to the elevated engagement.

Whether this change in engagement over time differs between conditions is still an open question, especially with regard to the tablet-only condition. When looking at the difference between the two robot conditions, it seems that task engagement decreases faster in the iconic gestures condition. It is possible that especially in this condition the robot distracted the children. As discussed earlier, and elaborated upon in D6.3, the gestures were performed slowly by the robot and there were many, perhaps too much, repetitions of each gesture. This might have led to a lower task engagement for the relatively fast 5-year-olds. Moreover, the trend seen for the tablet-only condition suggests that the novelty effect might not be specific to the robot, but could be a general effect of technology assisted learning. However, as we only had data from two lessons for the tablet-only condition, this conclusion is still premature.

To conclude, this preliminary engagement analysis offers some first insights into the engagement data and suggest that there is a decline in both forms of engagement over time. In addition, it might be that the presence of the robot and its use of (iconic) gestures have distracted children from the task rather than having aided them in learning. The preliminary data of the two robot conditions suggest that children were focused more on the robot and less on the task in the condition with iconic gestures than in the condition without iconic gestures. However, future analyses with data from the total sample are needed to be able to give more firm answers and conclusions.

4 Integrative discussion and recommendations

This deliverable reports on further analyses we conducted on the data of the large scale L2TOR evaluation study. The main outcomes of the large-scale study were that children learned from our training programme, that children did not learn more when the robot was present in addition to the tablet during the training, and that the robot's gestures did not benefit learning. The analyses on individual differences discussed in this deliverable showed differential effects the language and attentional skills. L1 vocabulary and phonological memory benefited learning in the robot without iconic gestures condition, while selective attention particularly benefited learning in the iconic gestures condition. Children with larger L1 vocabularies and/or better phonological memory benefited from the robot's presence, while not needing its iconic gestures to learn the target words. The results on selective attention suggest that good attentional skills are required to benefit from the robot's gestures. Our preliminary engagement data suggest that there is a slight decline in both task and robot engagement over time, and that the robot and its gestures may have distracted children from the task rather than having aided them in learning. Future analyses with data from the total sample will allow us to draw more firm conclusions regarding our research questions.

The main results from the large-scale study were already discussed in D7.2. The most important finding is that children did not learn *more* in the robot conditions than in the tablet-only condition. As discussed in D7.2 and D7.5, these findings are probably due to the strong focus on the tablet during the lessons. The tablet was required given the current state of technology, but led to an interaction that revolved more around the tablet than around the robot. As the robot cannot yet understand children's speech and cannot detect objects, interactions that do justice to the potential of robots could not yet be designed. The robot's physical presence and its possibilities for play with children are robots' most important advantage over other forms of technology such as tablets. These advantages currently mostly exist in theory but cannot be implemented in practice yet. Note, however, that the results may differ for older age groups, as speech recognition works slightly better for older children, and these children may respond differently to the robot. See also D7.5 for a more elaborate discussion of the lack of benefits from the robot in our study, and the paper by de Wit et al. (in preparation) in D6.3 for a discussion on the use of mediating devices such as tablets in human-robot interaction. We would like to emphasize that our study was one of the first longitudinal, preregistered, and sufficiently powered experiments conducted within HRI, and that even though the results may not support the current implementation of robots in language education, they do provide important insights into HRI.

It is clear that there are further technological developments necessary before robots can be used to support language learning in a way that does justice to their potential. For example, speech recognition and object recognition are needed to develop interactions in which robots can to some extent understand children and play with physical objects. With such developments, lessons can be developed that make use of the possible advantages of the robot, rather than working around the technical limitations of the robot. To work around the technical limitations, we designed a system in which the robot was very static. It followed predefined scripts, in which researchers and developers had invested many hours to develop. This also means that the robot is not flexible: it cannot divert from its

script depending on the situation. For example, the robot does not change the way it explains the meaning of a certain word based on the responses and actions of the child. If a child does not provide the correct response, a predefined set of feedback is provided. However, the robot cannot tell when children are distracted or need a different way of tutoring to understand the material. It cannot sense this input, and even if it could, it would not 'know' how to deal with it. Within our project, researchers have focused on trying to make the robot adaptive, that is, to try to adjust the difficulty of the lesson to the learner's knowledge using sophisticated Bayesian models (Schodde, Bergmann, & Kopp, 2017; De Wit et al., 2018). However, even adaptive robots appeared to be limited in their possibilities. Robots cannot, given the current state of technology, adapt depending on all relevant behaviours: they cannot yet monitor simultaneously the learner's knowledge, mental state, emotions, and movements, and adapt accordingly.

Having said all this, we do believe that robots have potential and we expect robots to become part of the educational landscape in years to come, although perhaps in a different way. We would like to present our view on how robots can, in the future, be implemented in educational contexts. Perhaps robots need to be much more intelligent to truly harbour their potential. There have been large developments in artificial intelligence in recent years, and robots until now rarely incorporate the most advanced artificial-intelligence systems. In their seminal paper, Smith and Gasser (2005) discuss six lessons learned from the development of human infants that should, in their view, guide the development of embodied intelligent agents (usually taken to imply AI systems). Perhaps robots need to go beyond being a physical body with simple computers in it to entities with artificial intelligence-systems that have a sort of embodied intelligence. The six lessons Smith and Gasser drew from babies are the following, in short: be multimodal (i.e., have concepts that are intrinsically grounded in and defined by coordinated multiple sensory and action schemes), be incremental (i.e., learn), be physical and explore (i.e., learn about the real world in real time), be social (i.e., be empathetic and learn about social rules), and learn a language (which should not be only about word-word relations, but also about word-world relations; cf. Pulvermüller, 2013). For the remainder of this section, we will assume that it is possible to develop an embodied intelligent agent, at least to some extent, according to these recommendations, and that a robot would incorporate such a system. Below, we discuss each of these six lessons and describe how a robot as a language tutor would benefit from being an embodied intelligent agent. It seems clear that not all recommendations can be equally easily followed-up due to hardware constraints and other technological issues.

The first lesson concerns multimodality: children learn through the various ways in which they come into contact with the environment, such as vision, audition, touch, and smell. They learn that their sensory systems are interrelated and the primary concepts they develop about the world consist in coordinated multimodal sensorimotor schemes. For example, the perception of an object changes if it is grabbed and moved, while at the same time the time-locked coordination of the varying perceptions with the motor movements underlie the integrated perception of invariant structure, which is the basis of multimodal object knowledge in the human infant. In our current robot, the robot uses few sensory systems and the different systems are not truly interrelated. Moreover the knowledge of the robot is essentially amodal and abstract (e.g. visual input is translated into a general information format, loosing much of its modality-specific richness). The robot in our

experiments received input from only the tablet and its cameras (which it can only use for face tracking and not for other types of vision such as object recognition). A robot that would have multiple sensory systems which it could integrate and relate to movement information in real time, would be a very different robot tutor. This robot would be able to perceive objects as invariant structures despite the ever changing perceptions when manipulating objects, it would create concepts which are grounded in real-life experiences with objects, and it would be able to perceive and act-upon objects as they are presented in a real-time situation. As a result, its gestures would also be grounded in its experiences with objects. The current way of developing robot gestures is a time-consuming procedure of modelling gestures after how one, the programmer, thinks a gesture should look like. However, gestures may be much more subtle and grounded in one's own experiences. A robot that would have held a heavy object, could subsequently gesture "holding" or "heavy" according to its own experiences with holding heavy objects.

The second lesson concerns incremental learning. Children's vision and motor development are related to their cognitive development such that their vision and motor abilities match and promote cognitive development. Currently, the robot is quite static and not learning. First steps are made towards adaptive robots, as discussed above, but the extent to which robots are really incremental is limited. An incremental robot tutor would adapt the difficulty of its lessons based on what it has learned from the child and the child's current needs in the concrete instruction situation. A beginner learner may need a "simpler" robot, which does not display too many complex social behaviours, than a more advanced learner. The robot can incrementally add behaviours as the learner progresses. This would also likely counter the novelty effect – the observed decline in motivation and interest of the child in interacting with the robot. Previous research has found that a robot that adds new behaviours over time results in child-robot interactions of higher and enduring quality than predictable robots without new behaviours (Tanaka, Cicourel, & Movellan, 2007). An incremental robot is less prone to children losing interest in the robot after having played with it for a longer period of time.

Lesson three is to be physical. Infants learn through interacting with physical objects and by linking objects, locations, and space. They can even learn words for objects that are not visible anymore while being labelled, simply by linking the label to the location in which the object was visible initially. The robot cannot really interact with the environment. It can move itself through space, but it does not perceive the environment while moving and has no spatial representation of its actions and of the perspective of the interaction partner. The robot in our study could manipulate the tablet, not by physically manipulating it externally, but through internal codes that moved objects on the tablet while the robot was moving its arm. An embodied physical robot would be able to use objects in its lessons. It would be able to recognize and hold objects, and thus to engage in lessons in which the focus lies not on the materials (as was the case in our study, due to the tablet) but on the robot and the child interacting with these materials.

Lesson four is to explore. Infants learn by engaging in actions with no apparent goal. Such actions help them to learn, amongst others, about action-consequence sequences and about the affordances of objects in a particular spatial lay-out, also uncommon affordances for action leading to new uses of objects (e.g. as tools). Children's exploration can be regarded as very rapid real-time learning about objects and what they afford in a given situation, which underlies adaptivity and creativity. Our current robot cannot respond

to events which are not preprogramed in the script, and cannot change its lesson and instruction behaviour depending on new, not pre-programmed events or object structures in the environment. Exploration in the sense of rapid real-time learning of action possibilities may be necessary for a robot to become truly adaptive. It can perceive the environment, draw the learner's attention to relevant or new stimuli in the environment, and respond meaningfully to unexpected events.

The fifth lesson is to be social and this may be the most difficult challenge. Infants learn social behavior through imitating their parents, and the parents provide social information (such as facial expressions and vocalizations) which the infant can imitate, matching the infant's developmental stage. However, it is not merely about imitation, or 'echoing', social and emotional cues expressed by the face, body posture, movement patterns of others (a challenge which could be in principle technically mastered by robots in due time). It is also about the direct coupling of this echoing, mimicking and imitation of others' behaviour to the child's own emotion systems (Gallese & Cuccio, 2015), enabling what philosophers of mind call direct access to the feeling states of others, underlying empathy and sympathy, giving motivational power to social (rule-following) and moral behavior (Tomasello & Vaish, 2013). Parents also tie action to sound, by matching their speech to the specific action that takes place (e.g., putting emphasis on a verb while showing the motion). An embodied intelligent agent can adapt its social behavior to the child's needs. It also can couple action and sounds, which can only be done manually – that is, through human interpretation and empathy - in our current robot. The robot's gestures in our study had to be time-locked to its speech by us, and thus may have differed from how humans would combine language and gestures naturally.

The sixth and last lesson is to learn a language, which is another challenging task. Language is a symbol system, in which sounds are arbitrarily mapped onto meaning. Language-in-use is also a system to share meaning through arbitrary but understood symbols that refer to the real world. Language as a symbol system can be abstracted from the real world, disregarding the referential meaning of language. Language, in this sense, is a computational system of word-word relations, but its connection to real world state of affairs, actions and events is problematic (Pulvermüller, 2013). Robots place us for a challenging question: what is true language comprehension and use? The current robot can speak, but cannot be said to have any comprehension of its utterances in terms of word-world relations. The sounds it produces are, referentially, as meaningless to the robot as any other sound. The robot can detect sound and convert speech streams from adults into text which it can subsequently use to respond, but it still does not have a comprehension of the adult's speech. Recognition of child speech is still a hurdle hard to take. Although it can be expected that this hurdle can be overcome in due time, this still begs the question if the robot indeed understands what a child is saying. Some natural-language processing and generation systems have been developed much further and can receive and produce speech without developers scripting each and every answer beforehand. However, do such systems truly have language? They do not have their concepts grounded in physical interactions with the environment nor in empathy-based social interactions with others, and perhaps such interactions are needed to truly comprehend and use language with all its subtle meanings. An embodied intelligent robot agent that would have similar concepts as language would engage in very different interactions than our current robot. For example,

it could use child-directed speech, interpret the child's current understanding and intentions, and use its knowledge grounded in interactions with the environment to gesture.

These six lessons illustrate that many technological developments are needed before it would be possible to develop an embodied intelligent robot agent that could deploy a robot's full potential in educational situations. Some of these technological developments are already nearby, others will take more time. And some other requirements may be impossible to meet. Apart from the question whether it is possible to develop robots in such a way, however, the question arises whether it is *desirable* to develop robots in such a way. In the most optimistic scenario, such an embodied intelligent robot agent would be capable of imitating human teachers and likely to be a very effective teacher. However, this is only true if children actually respond to a robot tutor the way they respond to a human tutor. This is still a relatively unexplored area of research. We have taken a first step in this direction with our study on children following a robot's *versus* a human's non-verbal behavior (described in D7.4). This study showed that children relied on a robot's non-verbal behavior similarly to that of a human. Moreover, our study on children's anthropomorphism of robots (described in D7.2) showed that many children have a tendency to attribute human-like characteristics to robots, such as having mental states and high-functioning cognitive abilities. At least some children appear to treat robots very similarly to humans. This raises fundamental ethical questions. Given that children seem to perceive robots as very similar to humans, is it ethical to develop robots that are highly similar to humans? Robots do not have the ability to truly understand feelings, nor do they have a moral compass or empathy-based motivation to care for children, while their highly human-like behavior may lead people, especially young children, to believe that they do. In a very basic sense we are deliberately deceiving the children. The ethics of developing robots for education should, therefore, be given a much more central place in the field.

Taking all together, it is questionable whether it will be possible to develop a robot according to these six lessons, and to develop a robot that can be multimodal, incremental, physical, social, explore, and master language. Without concluding that is not possible, it seems certain that the required technological developments demand huge investments and the question is whether these investments will ultimately pay-off. Perhaps it is more worthwhile to take a different approach to developing robots for education. Instead of trying to develop robots that can copy human tutors, we should look for ways in which robots can *complement* humans. If a robot is designed to copy a human, it will inevitably fall short of people's expectations sooner or later, at least given the current state of technology. Robots simply cannot behave exactly like humans. Rather, we should look how the different type of intelligence robots have can be used in an optimal way. For example, compared to humans, robots have infinite patience, do not get bored, can be designed specifically to serve, have a virtually unlimited memory capacity, have computational power, and a potentially unlimited repository of knowledge (e.g., through connections with the internet). And, indeed, robots can 'speak' and 'recognize' multiple languages, carry multiple languages' grammars, dictionaries and stories, as was one of the current project's starting points. Such qualities are especially valuable for "simple" tasks in which the learner needs extensive practice but does not need the robot to have highly interactional qualities. For example, a robot can help children learn tables and solve mathematical equations, or it can supervise independent seat work or motivate children to do their homework. And, indeed, a robot can help language learning children by providing

them with the dictionary items or grammatical examples needed in particular 'simple' language learning tasks, like learning a second language vocabulary or translating words. In such tasks, the robot's function is clear and does not mislead children by appearing much more communicative and socially skilled than it actually is. There are many situations in which robots can have a contribution to education, and in which we can clearly manage children's expectations beforehand to make sure that the robot will not fall short of them.

Another area to pursue in incorporating robots in education is related to the ethical question we raised. Educational programmes might focus on teaching children to 'understand' robots. Children can be taught how to interact with robots, what robots can and cannot do, what impression they may evoke but to what extent these are true and false, and so forth. Robots are, in many respects, a new species that we still do not understand well, unlike other nonhuman species that populate our (domestic) environments since ages. Equipping children with such knowledge about technologies such as robotics and AI will enable them to function in a more conscious and critical manner in a world where these technologies are increasingly incorporated in our daily lives.

In short, our project has shown that there are many technological limitations that have prevented us from designing robot-assisted lessons that can truly use a robot's potential for second language learning. The project has also shown that children can enjoy learning with the robot, and that there is potential for areas such as feedback and adaptivity to enhance learning. In the future, robots should be developed in a way which makes them much more embodied than they are now, following the six lessons from Smith and Gasser (2005). However, certain limitations seem impossible to overcome. Moreover, rather than trying to make robots as similar to humans as possible, perhaps we should focus on investigating whether there are ways in which robots can *complement* humans. Despite the limitations and limited effectivity found in the current study we do believe that social robots have potential added value for educational contexts. However, further technological advances are required, as well as better understanding of how children perceive and interact with robots. Moreover, we need to re-think the way we employ social robots in educational contexts in order for them to offer real advantages and effective educational interventions.

Reference list

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158–173. doi:10.1037/0033-295X.105.1.158

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2015). A quasi-universal nonword repetition task as a diagnostic tool for bilingual children learning Dutch as a second language. *Journal of Speech, Language, and Hearing Research*, *58*(6), 1747-1760.

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., ... & Vogt, P. (2018, February). 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* (pp. 50-58). ACM.

de Wit, J., Pijpers, L., van den Berghe, R., Krahmer, E., & Vogt, P. (in preparation). Why UX research matters for HRI: The case of tablets as mediators.

Dunn, L. M., Dunn, L. M., & Schlichting, L. (2005). *Peabody picture vocabulary test-III-NL*. Amsterdam: Pearson.

Gallese, V., & Cuccio, V. (2015). The paradigmatic body: Embodied simulation, intersubjectivity, the bodily self, and language. In T. Metzinger & J.M. Windt (Eds.), *Open Mind*: 14(T) (pp. 1-22). Frankfurt a. Main: Open Mind Group. doi: 10.15502/9783958570269.

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*, 513–543.

Gathercole, S. E., & Baddeley, A. D. (1990). The role of phonological memory in vocabulary acquisition: A study of young children learning new names. *British Journal of Psychology*, *81*, 439–454.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*, 1368–1378.

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

Kanero, J., Geçkin, V., Oranç, C., Mamus, E., Küntay, A. C., & Göksun, T. (2018). Social Robots for Early Language Learning: Current Evidence and Future Directions. *Child Development Perspectives, 12,* 146-151.

Konishi, H., Kanero, J., Freeman, M. R., Golinkoff, R. M., & Hirsh- Pasek, K. (2014). Six principles of language development: Implica- tions for second language learners. Developmental Neuropsychology, 39, 404–420.

Laevers, F., Daems, M., De Bruyckere, G., Declercq, B., Silkens, K., Snoeck, G., ... & Van Kessel, M. (2005). Zelfevaluatie-instrument voor welbevinden en betrokkenheid van kinderen in de opvang (ZiKo).

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

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), 422–429. doi:10.1080/09658210344000323

Mulder, H., Hoofs, H., Verhagen, J., 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. doi:10.3389/fpsyg.2014.00733

Pulvermüller, F. (2013). Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits. *Brain & Language*, *127*, 86-103.

R Core team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/.

Rowe, M. L. (2013). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development*, *83*, 1762–1774. doi:10.1111/j.1467-8624.2012.01805.x.A

Schmidt, R. W. (1990). The role of consciousness in second language. *Applied Linguistics*, *11*, 129–158.

Schodde, T., Bergmann, K., & Kopp, S. (2017, March). Adaptive robot language tutoring based on bayesian knowledge tracing and predictive decision-making. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 128-136). Chicago: ACM.

Service, E. (1992). Phonology, working memory, and foreign-language learning. *The Quarterly Journal of Experimental Psychology*, *45A*, 21–50.

Smith, L., & Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial life*, *11*, 13-29.

Stevens, C., & Bavelier, D. (2012). The role of selective attention on academic foundations: A cognitive neuroscience perspective. *Developmental Cognitive Neuroscience*, *2S*, S30–S48. doi:10.1016/j.dcn.2011.11.001

Tanaka, F., Cicourel, A., & Movellan, J. R. (2007). Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences*, *104*(46), 17954-17958.

Tomasello, M., & Vaish, A. (2013). The origins of human cooperation and morality. *Annual Review of Psychology*, *64*, 231-255.

van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., van der Ven, S., & Leseman, P. (2018). Social robots for language learning: A review. *Review of Educational Research.*

Verhagen, J., & Leseman, P. (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

Zaga, C., Lohse, M., Truong, K. P., & Evers, V. (2015). The Effect of a Robot's Social Character on Children's Task Engagement: Peer Versus Tutor. *International Conference on Social Robotics*, 704-713.

Appendix A: Coding scheme for engagement (in Dutch)

Protocol L2TOR coderen engagement

Deze handleiding is gebaseerd op een uitgebreid getest meetinstrument genoemd “ziko”. Voordat je aan de slag kan, moet je 1 van de begrippen achter het instrument leren kennen: *engagement*.

Het is belangrijk dat je vooraf leert om gericht te kijken naar kinderen en weet hoe je moet werken
met het instrument. Enkel als je de handleiding onder de knie hebt, kan je de scores juist invullen. De
voorbereiding van de zelfevaluatie is van groot belang. Wil je meer informatie en hulp bij het inoefenen van ziko? Daarvoor kan je terecht bij ecego. Je vindt meer informatie op de website van kind en gezin (www.kindengezin.be) en van ecego (www.cego.be)

Wat is engagement

Een kind dat engaged is, wordt als het ware ‘helemaal opgeslorpt´ in zijn activiteit:
Spelen met blokken, boetseren of puzzelen, luisteren naar een verhaal, met anderen praten, het
is een heel aparte beleving die je zowel bij baby´s als bij volwassenen kan herkennen.

Motivatie

Als je engaged bent, voel je je aangesproken door de activiteit, dus ben je werkelijk
geïnteresseerd. Engagement krijg je niet als je dingen alleen maar doet omdat anderen
het vragen of er jou toe verplichten. Je motivatie komt vanuit jezelf, dit kan dus wel opgedragen zijn vanuit anderen, maar je bent er zelf actief mee bezig.

Intense mentale activiteit

Bij engagement stel je je helemaal open voor ervaringen: de indrukken die je opdoet zijn heel sterk. Lichaamsgewaarwordingen en bewegingservaringen, kleuren en klanken, geuren en smaken hebben een schakering en een diepte die er anders niet zijn. Je spreekt je verbeelding en je denkvermogen ten volle aan. Bij niet-betrokken activiteit zijn de gewaarwordingen niet doorleefd, dus oppervlakkig.

Voldoening

Engagement is een heerlijke toestand: je bent in vervoering. Wat je beleeft is energie die door je stroomt. Kinderen nemen spontaan steeds opnieuw initiatieven die hen in die toestand brengen. Spel is de plek bij uitstek waarin ze deze genoegdoening vinden.
Ontbreekt engagement, dan krijg je verveling, een gevoel van leegte en frustratie.

Exploratiedrang

De bron voor engagement is de ontdekkings- of exploratiedrang, de drang om de wereld te ervaren, om zintuiglijke indrukken op te doen, om greep te krijgen op de werkelijkheid. Aanvankelijk is dat 'greep krijgen´ letterlijk te nemen: aanraken en grijpen wat in de buurt komt. Gaandeweg gaat het meer om het 'begrijpen´ van de werkelijkheid.

Aan de grens van je mogelijkheden

Engagement is mogelijk als een activiteit een uitdaging is, niet te makkelijk en ook niet te moeilijk. Bij engagement bewegen mensen zich dus aan de grens van hun mogelijkheden. Ze spreken hun vermogens ten volle aan, ze geven het beste van zichzelf - of we het nu over baby´s hebben of volwassenen, over kinderen met een zwakke mentale ontwikkeling of over hoogbegaafden.

Waar engagement goed voor is

Engagement is iets heel bijzonders. Iedereen die gewoon naar kinderen kijkt, wordt erdoor verrast. Je voelt intuïtief aan dat je het spel niet mag verstoren.
Is er engagement, dan weten we dat kinderen hun mogelijkheden aanspreken en dat ze 'in ontwikkeling´ zijn: ze leren op een dieper niveau, ze worden echt competenter.

Jouw taak

Je gaat de engagement van het kind bepalen. Je werkwijze is eenvoudig en te vergelijken met 'scannen´: je observeert het kind gedurende een tweetal minuten (1 videofragment). Geef elk kind een score voor engagement op basis van een vijfpuntenschaal. Je mag ook halve punten geven, dus het kind kan ook 3.5 engaged zijn. Bij het scannen gaat het om een momentopname, het kan dus zijn dat hetzelfde kind het ene fragment een lage engagement scoort en het andere moment een hogere engagement. Daarnaast kijk je naar de engagement over het gehele fragment. Laat het kind dus in het begin van het fragment een hogere engagement zien dan in het laatste gedeelte; dan middel je over deze twee waardes. Dit middelen laat je ook afhangen van de periode dat het kind deze engagement

laat zien, als het kind dus een derde van het videofragment een hoge engagement (5) laat zien en gedurende 2 derde van het fragment een lagere engagement (3) laat zien. Dan is de uiteindelijke niveau voor engagement dus tussen een 3.5 en een 4. Handig is dus om tijdens de fragmenten te noteren hoe engaged het kind is en waarom je dat vindt.

We gaan twee soorten engagement meten: child-task engagement en child-robot engagement.

Het meten van child-task engagement

Child-task engagement kijkt naar hoe de kind engaged is met de taak. Dit kan op de tablet zijn, maar ook als de robot vraagt dat het kind iets moet doen (zoals nazeggen en nadoen). Als het kind doordat de robot praat richting de robot kijkt, is het kind nog steeds engaged met de taak. Ook in het geval dat het kind naar de robot kijkt als de robot een gebaar laat zien, leidt dit niet tot een lagere child-task engagement. Immers, het nazeggen en de gebaren behoren tot de taak. Alleen in het geval dat het kind ergens anders op focust tijdens de taak of naar de robot kijkt zonder enige reden scoor je de child-task engagement lager. Dit betekent ook dat je niet meet hoe engagement het kind met de robot is, dat is de focus van de andere engagement schaal.

Child-task engagement gaat ook gepaard met fouten in het spel, over het algemeen leidt een lagere engagement tot meer fouten bij een kind. Maar, zoals jullie vast herkennen door het zelf afnemen van de experimenten, zag het systeem soms fouten die eigenlijk niet fout waren. In dit geval is het aan jou om deze fouten niet mee te laten tellen met jouw child-engagement score.

Hieronder de schaal voor child-task engagement in een tabel met voorbeelden gezet.

De schaal voor child-task engagement

Niveau	Engagement	Voorbeelden
1	Uitgesproken laag	Het kind vertoont nagenoeg geen activiteit: „ Geen concentratie: staren, wegdromen; „ Een afwezige, passieve houding; „ Geen gerichte activiteit, doelloze handelingen, niets teweegbrengen; „ Alleen bezig met de experiment leider en niet met de taak; „ Geen tekenen van exploratie en interesse; „ Niets in zich opnemen, geen mentale activiteit.

2	Laag	Het kind vertoont enige activiteit, maar deze wordt geregeld onderbroken: „ Beperkte concentratie: wegkijken, prullen (friemelen), dromen; „ Makkelijk afgeleid; „ Taken worden in beperke mate uitgevoerd.
3	Matig	Er is de hele tijd activiteit, maar niet echt geconcentreerd. „ Het kind is routinematig, vluchtig bezig; „ Is beperkt gemotiveerd, voelt zich niet uitgedaagd, toont geen echte inzet; „ Doet geen diepgaande ervaring op; „ Is niet opgeslorpt door wat het doet; „ Gebruikt zijn capaciteiten maar met mate; „ De activiteit raakt de verbeelding en het denkvermogen van het kind niet. „ De meeste taken worden uitgevoerd.
4	Hoog	Er zijn doorgaans signalen van engagement: „ Het kind gaat globaal op in zijn spel; „ Er is doorgaans concentratie, maar soms verslapt de aandacht; „ Het kind voelt zich uitgedaagd, er is een zekere gedrevenheid; „ Gebruikt zijn capaciteiten; „ Spreekt de verbeelding en het denkvermogen aan.

5	Uitgesproken hoog	Het kind is gedurende de hele tijd ononderbroken bezig en gaat sterk op in zijn activiteit: „ Is ononderbroken geconcentreerd, opgeslorpt door de activiteit, vergeet de tijd; „ Is heel gemotiveerd, voelt zich sterk aangesproken; „ Is niet af te leiden; „ Kijkt aandachtig naar de taak, heeft aandacht voor details; „ Spreekt voortdurend al zijn capaciteiten en mogelijkheden aan; „ Er is een sterke mentale activiteit: de verbeelding en het denkvermogen draaien op volle toeren; „ Doet diepgaande nieuwe ervaringen op; „ Geniet van zo gedreven bezig te zijn.

Het meten van child-robot engagement

Child-robot engagement kijkt alleen naar hoe de kind engaged is met de robot. Dit is niet gerelateerd aan de taak. Het kind kan engaged met de robot zijn zonder dat het kind de taak uitvoert. Child-robot engagement wordt bepaald door de mate van hoe vaak het kind praat met de robot en kijkt richting de robot. Alleen het nazeggen van een target word is geen teken van child-robot engagement, immers de kinderen in de tablet conditie praten ook de tablet na. Als het kind bij het nazeggen van het target woord de robot ook nog aankijkt, dan telt het wel mee voor de child-robot engagement. Ook kinderen die de gebaren van de robot na doen laten een hoge engagement zien. Een kind dat alleen richting de tablet kijkt en de robot negeert (probeert te negeren) zal juist lager scoren.

De schaal voor child-robot engagement		
Niveau	Engagement	Voorbeelden

1	Uitgesproken laag	Het kind vertoont nagenoeg geen interactie: „ Geen concentratie: staren, wegdromen; „ Negeert de robot volledig; „ Heeft een gesloten (lichaams)houding richting de robot; „ Een afwezige, passieve houding; „ Geen gerichte activiteit, doelloze handelingen, niets teweegbrengen; „ Geen tekenen van exploratie en interesse; „ Niets in zich opnemen, geen mentale activiteit.
2	Laag	Het kind vertoont enige interactie, maar deze wordt geregeld onderbroken: „ Beperkte concentratie: wegkijken, prullen (friemelen), dromen; „ Kijkt beperkt richting de robot; „ Makkelijk afgeleid; „ Handelingen leiden maar tot beperkt resultaat.
3	Matig	Er is de hele tijd activiteit, maar niet echt geconcentreerd. „ Het kind is routinematig, vluchtig bezig; „ Is beperkt gemotiveerd, voelt zich niet uitgedaagd, toont geen echte inzet; „ heeft een open (lichaams)houding richting de robot; „ Doet geen diepgaande ervaring op; „ Is niet opgeslorpt door wat het doet; „ Gebruikt zijn capaciteiten maar met mate; „ De activiteit raakt de verbeelding en het denkvermogen van het kind niet. „ Doelloos aanraken van de robot
4	Hoog	Er zijn doorgaans signalen van engagement: „ Het kind gaat globaal op in zijn spel met de robot; „ Er is doorgaans sprake van joint attention; „ Er is doorgaans concentratie, maar soms verslapt de aandacht; „ Het kind voelt zich uitgedaagd, er is een zekere gedrevenheid; „ Gebruikt zijn capaciteiten; „ Spreekt de verbeelding en het denkvermogen aan.

5	Uitgesproken hoog	Het kind is gedurende de hele tijd ononderbroken bezig en gaat sterk op in zijn activiteit met de robot: „ Is ononderbroken geconcentreerd, vergeet de tijd; „ Is heel gemotiveerd, voelt zich sterk aangesproken; „ Is niet af te leiden; „ Kijkt aandachtig naar robot, heeft aandacht voor details; „ Praat tegen de robot; „ Gebaren na doen (alleen in de iconische gebaren conditie); „ Er is sprake van joint attention; „ Er is een sterke mentale activiteit: de verbeelding en het denkvermogen draaien op volle toeren; „ Doet diepgaande nieuwe ervaringen op; „ Geniet van zo gedreven bezig te zijn.

Appendix B: Results of analyses on individual differences

Below, the tables for each of the models and each of the language and attentional skills (i.e., L1 vocabulary, phonological memory, and selective attention) can be found. The "B" is an indicator of the effect size, and the *p*-value indicates significance.

Results for L1 vocabulary.

Table B.1.
Results from the generalized linear regression model with scores from the English-Dutch translation task as a dependent variable, condition and L1 vocabulary as between-participants fixed effects, and time as a within-participants fixed effect. Significant effects are boldfaced.

	B	**SE**	***Z***	***p***
Condition contrast 1	**1.14**	**0.30**	**3.77**	**< 0.001**
Condition contrast 2	0.31	0.26	1.21	0.228
Condition contrast 3	-0.02	0.30	-0.06	0.951
Time contrast 1	**-1.37**	**0.09**	**-15.00**	**< 0.001**
Time contrast 2	**0.16**	**0.05**	**2.97**	**0.003**
L1 vocabulary	**1.83**	**0.92**	**1.99**	**0.046**
Condition contrast 1 * time contrast 1	**-1.02**	**0.23**	**-4.52**	**< 0.001**
Condition contrast 2 * time contrast 1	0.19	0.18	1.05	0.292
Condition contrast 3 * time contrast 1	0.07	0.22	0.31	0.755
Condition contrast 1 * time contrast 2	-0.11	0.14	-0.81	0.419
Condition contrast 2 * time contrast 2	0.04	0.10	0.40	0.688
Condition contrast 3 * time contrast 2	-0.21	0.12	-1.77	0.077
Condition contrast 1 * L1 vocabulary	**4.77**	**2.17**	**2.20**	**0.028**
Condition contrast 2 * L1 vocabulary	-3.69	2.14	-1.72	0.085
Condition contrast 3 * L1 vocabulary	2.46	2.39	1.03	0.304

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures. Time contrast 1: post-tests vs. pre-test. Time contrast 2: delayed vs. immediate post-test.

Table B.2.
Results from the generalized linear regression model with scores from the English-Dutch translation task and Dutch-English translation task as dependent variables, condition and L1 vocabulary as between-participants fixed effects, and time and language as within-participants fixed effects. Significant effects are boldfaced.

	B	**SE**	***Z***	***p***
Condition contrast 1	**1.86**	**0.37**	**4.99**	**< 0.001**
Condition contrast 2	0.20	0.28	0.70	0.483
Condition contrast 3	-0.10	0.32	-0.31	0.757
Time	**0.16**	**0.05**	**3.28**	**0.001**
Language	**-0.56**	**0.05**	**-11.11**	**< 0.001**
L1 vocabulary	**25.19**	**8.96**	**2.81**	**0.005**
Condition contrast 1 * time	-0.15	0.15	-1.03	0.302
Condition contrast 2 * time	0.05	0.10	0.47	0.640
Condition contrast 3 * time	-0.20	0.12	-1.70	0.090
Condition contrast 1 * L1 vocabulary	**72.47**	**8.99**	**8.06**	**< 0.001**
Condition contrast 2 * L1 vocabulary	**-24.52**	**10.94**	**-2.24**	**0.025**
Condition contrast 3 * L1 vocabulary	**31.99**	**9.16**	**3.49**	**< 0.001**

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures.

Table B.3.
Results from the generalized linear regression model with scores from the comprehension task as a dependent variable, condition and L1 vocabulary as between-participants fixed effects, and time as a within-participants fixed effect. Significant effects are boldfaced.

	B	**SE**	***Z***	***p***
Condition contrast 1	**0.34**	**0.11**	**3.05**	**0.002**
Condition contrast 2	0.02	0.11	0.22	0.829
Condition contrast 3	-0.11	0.17	-0.66	0.513
Time	0.05	0.04	1.08	0.278
L1 vocabulary	4.01	4.13	0.97	0.332
Condition contrast 1 * time	-0.06	0.09	-0.60	0.550
Condition contrast 2 * time	0.04	0.09	0.44	0.657
Condition contrast 3 * time	-0.01	0.14	-0.05	0.958
Condition contrast 1 * L1 vocabulary	11.01	6.59	1.67	0.095
Condition contrast 2 * L1 vocabulary	-9.86	6.67	-1.48	0.140
Condition contrast 3 * L1 vocabulary	**19.56**	**6.43**	**3.05**	**0.002**

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures.

Results for phonological memory.

Table B.4.
Results from the generalized linear regression model with scores from the English-Dutch translation task as a dependent variable, condition and phonological memory as between-participants fixed effects, and time as a within-participants fixed effect. Significant effects are boldfaced.

	B	**SE**	***Z***	***p***
Condition contrast 1	**1.12**	**0.31**	**3.61**	**< 0.001**
Condition contrast 2	0.30	0.27	1.12	0.264
Condition contrast 3	-0.03	0.31	-0.11	0.911
Time contrast 1	**-1.37**	**0.09**	**-14.84**	**< 0.001**
Time contrast 2	**0.16**	**0.05**	**2.99**	**0.003**
Phonological memory	5.76	3.47	1.66	0.097
Condition contrast 1 * time contrast 1	**-1.02**	**0.23**	**-4.48**	**< 0.001**
Condition contrast 2 * time contrast 1	0.19	0.18	1.00	0.315
Condition contrast 3 * time contrast 1	0.07	0.22	0.32	0.750
Condition contrast 1 * time contrast 2	-0.12	0.14	-0.85	0.393
Condition contrast 2 * time contrast 2	0.05	0.10	0.47	0.637
Condition contrast 3 * time contrast 2	-0.19	0.12	-1.59	0.112
Condition contrast 1 * phonological memory	5.68	6.96	0.82	0.415
Condition contrast 2 * phonological memory	0.90	5.77	0.16	0.877
Condition contrast 3 * phonological memory	**15.79**	**6.67**	**2.37**	**0.018**

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures. Time contrast 1: post-tests vs. pre-test. Time contrast 2: delayed vs. immediate post-test.

Table B.5.
Results from the generalized linear regression model with scores from the English-Dutch translation task and Dutch-English translation task as dependent variables, condition and phonological memory as between-participants fixed effects, and time and language as within-participants fixed effects. Significant effects are boldfaced.

	B	**SE**	***Z***	***p***
Condition contrast 1	**1.82**	**0.38**	**4.80**	**< 0.001**
Condition contrast 2	0.12	0.29	0.40	0.688
Condition contrast 3	-0.10	0.33	-0.30	0.762
Time	**0.16**	**0.05**	**3.18**	**0.001**
Language	**-0.55**	**0.05**	**-10.93**	**< 0.001**
Phonological memory	**20.34**	**7.04**	**2.89**	**0.004**
Condition contrast 1 * time	-0.16	0.15	-1.08	0.282
Condition contrast 2 * time	0.06	0.10	-0.56	0.575
Condition contrast 3 * time	-0.18	0.12	-1.52	0.129
Condition contrast 1 * phonological memory	**22.13**	**8.07**	**2.74**	**0.006**
Condition contrast 2 * phonological memory	**26.72**	**10.53**	**2.54**	**0.011**
Condition contrast 3 * phonological memory	**40.52**	**14.99**	**2.70**	**0.007**

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures.

Table B.6.
Results from the generalized linear regression model with scores from the comprehension task as a dependent variable, condition and phonological memory as between-participants fixed effects, and time as a within-participants fixed effect. Significant effects are boldfaced.

	B	**SE**	***Z***	***p***
Condition contrast 1	**0.34**	**0.12**	**2.97**	**0.003**
Condition contrast 2	0.04	0.12	0.38	0.708
Condition contrast 3	-0.09	0.18	-0.49	0.621
Time	0.05	0.04	1.09	0.277
Phonological memory	0.32	4.13	0.08	0.939
Condition contrast 1 * time	-0.06	0.09	-0.60	0.552
Condition contrast 2 * time	0.04	0.09	0.44	0.659
Condition contrast 3 * time	-0.01	0.14	-0.06	0.956
Condition contrast 1 * phonological memory	3.21	5.89	0.55	0.586
Condition contrast 2 * phonological memory	-7.45	6.12	-1.22	0.224
Condition contrast 3 * phonological memory	**12.85**	**5.67**	**2.27**	**0.023**

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures.

Results for selective attention.

Table B.7.
Results from the generalized linear regression model with scores from the English-Dutch translation task as a dependent variable, condition and selective attention as between-participants fixed effects, and time as a within-participants fixed effect. Significant effects are boldfaced.

	B	**SE**	***Z***	***p***
Condition contrast 1	**1.21**	**0.31**	**3.95**	**< 0.001**
Condition contrast 2	0.27	0.26	1.01	0.312
Condition contrast 3	0.01	0.30	0.02	0.982
Time contrast 1	**-1.37**	**0.09**	**-14.91**	**< 0.001**
Time contrast 2	**0.16**	**0.05**	**3.13**	**0.002**
Selective attention	2.51	2.28	1.10	0.269
Condition contrast 1 * time contrast 1	**-1.02**	**0.23**	**-4.49**	**< 0.001**
Condition contrast 2 * time contrast 1	0.19	0.18	1.01	0.313
Condition contrast 3 * time contrast 1	0.09	0.22	0.40	0.691
Condition contrast 1 * time contrast 2	-0.11	0.14	-0.81	0.419
Condition contrast 2 * time contrast 2	0.04	0.10	0.37	0.709
Condition contrast 3 * time contrast 2	-0.21	0.12	-1.75	0.080
Condition contrast 1 * selective attention	**9.45**	**4.55**	**2.08**	**0.038**
Condition contrast 2 * selective attention	-4.28	4.21	-1.02	0.310
Condition contrast 3 * selective attention	**-10.22**	**4.98**	**-2.05**	**0.040**

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures. Time contrast 1: post-tests vs. pre-test. Time contrast 2: delayed vs. immediate post-test.

Table B.8.
Results from the generalized linear regression model with scores from the English-Dutch translation task and Dutch-English translation task as dependent variables, condition and selective attention as between-participants fixed effects, and time and language as within-participants fixed effects. Significant effects are boldfaced.

	B	**SE**	***Z***	***p***
Condition contrast 1	**1.90**	**0.38**	**5.04**	**< 0.001**
Condition contrast 2	0.13	0.29	0.46	0.649
Condition contrast 3	-0.09	0.33	-0.26	0.793
Time	**0.16**	**0.05**	**3.28**	**0.001**
Language	**-0.56**	**0.05**	**-11.11**	**< 0.001**
Selective attention	**20.58**	**6.66**	**3.09**	**0.002**
Condition contrast 1 * time	-0.15	0.15	-1.03	0.302
Condition contrast 2 * time	0.05	0.10	0.47	0.640
Condition contrast 3 * time	-0.20	0.12	-1.70	0.089
Condition contrast 1 * selective attention	**36.46**	**6.47**	**5.63**	**< 0.001**
Condition contrast 2 * selective attention	-13.53	10.42	-1.30	0.194
Condition contrast 3 * selective attention	**-41.25**	**8.75**	**-4.71**	**< 0.001**

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures.

Table B.9.
Results from the generalized linear regression model with scores from the comprehension task as a dependent variable, condition and selective attention as between-participants fixed effects, and time as a within-participants fixed effect. Significant effects are boldfaced.

	B	**SE**	***Z***	***p***
Condition contrast 1	**0.30**	**0.13**	**2.30**	**0.021**
Condition contrast 2	0.02	0.13	0.17	0.862
Condition contrast 3	-0.21	0.19	-1.10	0.272
Time	0.03	0.05	0.59	0.557
Selective attention	**9.06**	**4.61**	**1.96**	**0.050**
Condition contrast 1 * time	-0.01	0.12	-0.07	0.946
Condition contrast 2 * time	0.05	0.10	0.47	0.638
Condition contrast 3 * time	-0.02	0.15	-0.12	0.908
Condition contrast 1 * selective attention	10.40	8.07	1.29	0.197
Condition contrast 2 * selective attention	11.11	8.03	1.38	0.166
Condition contrast 3 * selective attention	-7.63	8.09	-0.94	0.346

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures.