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

D1.3 Human tutoring experiments

Due Date: **28/02/2018**
Submission Date: **31/03/2018**

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

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

Responsible Person: **Junko Kanero** Revision: **1.1**

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

The aim of this deliverable is to discuss the results of empirical studies with a human experimenter that examined effects of feedback types and of gesture types on second language learning. This document describes the design, procedure, data analyses, and results of the studies.

Principal Contributors

Koç Aylin Küntay, Tilbe Göksun, Junko Kanero, Cansu Oranç, Özlem Ece Demir-Lira, Idil Franko, and Sümeyye Koşkulu

Revision History

Version 1.0 (JK 31-03-2018)
First version.
Version 1.1 (JK 10-05-2018)
First version with corrections of typos and other minor errors.

1 Introduction

The notable strengths of robot second language (L2) tutors, as compared to other digital devices such as tablets, may be their abilities to be adaptive and to perform actions. In terms of adaptivity, robots can use different sensors to detect motivational and educational needs of learners and change their behavior accordingly. In other words, robots are capable of providing different types of *feedback* that can scaffold L2 learning in young children. Research in developmental psychology suggests that children rely on verbal feedback as well as implicit non-verbal feedback (e.g., eye gaze) to learn new words (e.g., Konishi, Kanero, Freeman, Golinkoff, & Hirsh-Pasek, 2014). Feedback has been used in educational robotics. For example, English-speaking 3- to 5-year-olds learned Spanish words successfully with a robot that provided explicit verbal feedback (e.g., "Good job!"), adjusting it based on the performance of students (Gordon et al., 2016). It can be difficult for classroom teachers to adjust lesson levels to each child. Robot tutors can serve as a supplementary tool, especially when children can practice one on one with the robot.

Humanoid robots can also perform actions and gestures in the real world. Exploring the use of robot gestures and other actions is especially important in L2TOR because research on interactions between adults and children suggests that (1) although the positive effects of gestures are found both for adults (e.g., Macedonia, Müller, & Friederici, 2011) and young children (e.g., Tellier, 2008), children benefit from gestures more than adults (Hostetter, 2011); (2) gestures improve L2 speech (Sueyoshi & Hardison, 2005); and (3) gestures increase children's attention to the learning materials (Valenzeno, Alibali, & Klatzky, 2003). Research in child-robot interaction also found gestures beneficial. For example, Italian-speaking 5- to 6-year-olds recalled stories more accurately when the tales were narrated by an expressive humanoid robot that used gestures, eye gaze, and voice tone than when they were told by an inexpressive human teacher (Conti, Di Nuovo, Cirasa, & Di Nuovo, 2017).

Although benefits of adaptive feedback and of gestures have been found, less is known about whether and how types of feedback and gestures make differences in language learning or learning in general. To understand the ways in which human teachers use feedback and gestures, we previously conducted semi-naturalistic observations at L2 classrooms (see Deliverable 1.2) and pilot studies with human teachers (Deliverable 1.1). Our observations suggest that human teachers use a wide variety of feedback and gestures to scaffold the learning process.

Feedback can be classified into several interrelated subcategories. First, there is positive feedback used to praise and encourage children (e.g., "Good job!") or to confirm children's response to a question (e.g., "Yes, that's it!") and negative feedback (e.g., "No, that's wrong."). Our observations revealed that human teachers use positive feedback significantly more often than negative feedback. At the same time, however, teachers must provide negative feedback when the response must be clearly corrected to ensure proper learning. Thus, when and how teachers should provide negative feedback is a critical issue. Second, positive and negative feedback can be further divided into implicit and explicit feedback (e.g., Carroll & Swain, 1993). For example, teachers can provide positive implicit feedback by repeating the child's correct response.

Gestures can also be classified into several categories. In teaching a word, a robot can perform *deictic gestures* to indicate the reference of a word, e.g., point to an object, or *iconic gestures* that represent the meaning of the word, e.g., opening its arms to represent the meaning of the word "big." In our semi-naturalistic observations, deictic gestures were abundant in L2 lessons. Iconic gestures, on the other hand, were used

much less though this must be at least partially because some of the taught words cannot be represented using iconic gestures (e.g., color names). In L2TOR, we specifically chose math and space as the two lesson domains because learning math and spatial words are challenging. Thus, children can benefit from extra help from gestures. The words in these two domains can also be gestured easily. For example, in the math domain, we teach adjectives that specify different features of objects such as size ("big" and "small") and height ("high" and "low"). Human adults have no problem with gesturing these words. Hence, it is important to explore the possibility of using not only deictic but also iconic gestures.

Our observations revealed that experienced human teachers use various kinds of feedback and gestures in teaching L2 to children. However, it is unknown whether the techniques used by human teachers are actually effective. Therefore, we conducted experimental studies with a human experimenter to identify types of feedback and gestures that are effective in teaching L2 vocabulary.

2 Overview of the Experimental Studies

To understand how different types of feedback and gestures influence L2 word learning in preschoolers, we conducted three experimental studies: Two studies examining feedback types and one study examining gesture types. The first feedback study evaluated effects of feedback types in learning of spatial prepositions (e.g., "above" and "on"), and the second study evaluated the same constructs in learning of motion verbs (e.g., "climb" and "slide"). For gestures, we originally planned to conduct two separate studies to look into effects of deictic gestures (Task 1.2) and of iconic gestures (Task 1.3). However, we decided to change the plan and test both types of gestures in one experiment so that these two gesture types can be directly compared. All experiments were conducted by KOC team in Turkey. In this deliverable, we report the results of the experiments in which a human experimenter interacted with children. Nevertheless, we have already created NAO programs to conduct robot versions of Verb Feedback Study and Gesture Study and plan to collect data for those studies in the coming months to examine whether feedback and gesture have different effects when they are given by a human experimenter or a robot.

3 Feedback Study

The two feedback studies we conducted focused on how to correct children to ensure proper word learning. In other words, we tested different types of negative feedback that can be given when the child makes a mistake. There are a wide variety of ways in which negative feedback can be given. We tested three different types of feedback – *repetition*, *description*, and *demonstration* – all of which are discussed in literature and observed in L2 classrooms (see Deliverable 1.2). Broadly speaking, *repetition* is implicit negative feedback in which the teacher simply repeats the prompt or question to inform that the child did not give the correct response. *Description* is explicit negative feedback in which the teacher describes and elaborates on the wrong response the child gave to show that the correct response is something else. *Demonstration* is explicit negative feedback in which the teacher herself demonstrates the correct response. The three types of negative feedback were used while children learned spatial prepositions (Preposition Feedback study) and motion verbs (Verb Feedback study). In addition, *negation* was tested in the Preposition Feedback study. In

the Negation condition, the child was explicitly told that their choice was not the correct one, but no additional hints were given.

3.1 Preposition Feedback Study

In the Preposition Feedback Study, a human experimenter used repetition, description, demonstration, and negation as feedback while teaching four spatial prepositional phrases: in front of, on, above, and under. The study also served as a pilot study for determining the design of Verb Feedback Study, which tested a larger number of participants.

3.1.1 Participants

Fourteen preschoolers participated in the Preposition Feedback Study (*Age range* = 47-78 months; *Mean age* = 58.64 months; *SD* = 10.25; 12 females). Participants had no known vision or hearing impairments. Five additional children were also recruited but were excluded from the study, as they did not finish the task.

3.1.2 Stimuli

The four target prepositions taught in this experiment were in front of, on, above, and under. Four images were created to represent each of the words (Figure 1). All images depicted a blue bird and an airplane. As shown in Figure 1, the bird was either sitting on the airplane, flying under the airplane, flying in front of the airplane, or flying above the airplane. In addition, we created three cards each of which depicting the same blue bird, a butterfly, or a plane. These additional cards were used just in the Introduction phase to teach the English nouns “bird” and “plane” to participants.

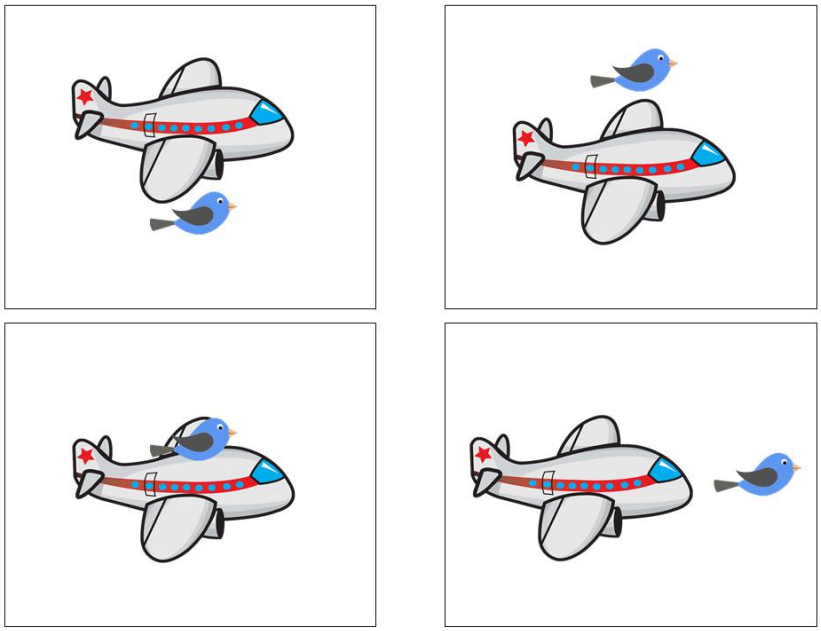

Figure 1. The images used to represent the four target prepositions taught in the Preposition Feedback Study –*under* (top left), *above* (top right), *on* (bottom left), and *in front of* (bottom right).

3.1.3 Design

The task was similar to the board game *Guess Who?* For children, the goal of the game was to guess the card the experimenter had in her hands. The experimenter

described her card in English (e.g., “The bird is on the plane”), and the child was asked to choose the target card that matches the experimenter’s description from four options shown on the tablet screen.

The study tested four types of feedback - simple repetition of the prompt (Repetition), negation of the child’s choice (Negation), description of the child’s choice (Description), and demonstration of the correct response (Demonstration; see Table 1 for examples). These types of feedback were among the most frequently observed in our English teacher observations in Turkish preschools (see Deliverable 1.2). In addition to the first three conditions that were proposed in Description of Action, Demonstration was added based on the classroom observations. Children were given the same feedback when their responses are correct (“Yes! The bird is on the plane”; “Yes! The bird is in front of the plane”; “Yes! The bird is above the plane”; or “Yes! The bird is under the plane”). All instructions, except the prompt (e.g., “This bird is on the plane”) and feedback (see Table 1), were given in the native language of participants, i.e., Turkish.

Table 1. Feedback provided by the experimenter in the Preposition Feedback Study. All feedback phrases were given in English except for “Tekrar dene” which means, “Try again” in English.

	Feedback to a correct response	Feedback to an incorrect response
Repetition	Yes! The bird is on the plane.	No! This bird is on the plane. This bird is on the plane. *Tekrar dene!*
Negation	Yes! The bird is on the plane.	No! That bird is not on the plane. This bird is on the plane. *Tekrar dene!*
Description	Yes! The bird is on the plane.	No! That bird is [under] the plane. This bird is on the plane. *Tekrar dene!*
Demonstration	Yes! The bird is on the plane.	No! This bird on the plane. (points to the correct image) Now you touch the bird on the plane. *Tekrar dene!*

There were four phases in the Preposition Feedback Study: Introduction, Learning, Practice, and Test. In the Introduction phase, the child was first told that they would play a game in which they were ought to guess the card which the experimenter had in her hand. Then, the child learned two English nouns that were used as the subject and object of the sentence (“bird” and “plane,” respectively), in which all target prepositions were embedded. The experimenter showed an image of a bird and an image of a plane, and described the meanings of the English words in Turkish. To ensure

children learned the two nouns, children were presented with three images on the tablet – a bird, a plane, and a butterfly – and were asked to touch "bird" and "plane."

In the Learning phase, the child was presented with the four images in Figure 1 one by one. For each image, the experimenter introduced a corresponding preposition by describing the spatial relation between the bird and the plane in English: "The bird is on the plane. *İngilizce'de on, üstünde demek*. On. The bird is on the plane. *Kuş uçağın üstünde. Şimdi benden sonra tekrar eder misin?* On. The bird is on the plane." (i.e., "The bird is on the plane. *In English*, on *means on*. On. The bird is on the plane. *The bird is on the plane*. *Now, can you repeat after me?* On. The bird is on the plane."). We introduced all target prepositions both in isolation (e.g., "on") and in a sentence (e.g., "the bird is on the plane") to ensure that children learn the meaning of the word as well as how it is used in sentences.

The Practice phase was essentially the Test phase but carried out in Turkish. The child was presented with the four images as shown in Figure 1 and was asked to choose an image that corresponds to a Turkish spatial relation word (postposition) – *altında* (under), *önünde* (in front of), *üzerinde* (above), or *üstünde* (on). This phase was included to ensure that the child understands the task and is able to complete the main task.

In the Test phase, the child was again shown four images of a bird and an airplane. They were the images used in the Practice phase except that the color of the bird changed across trials to keep children's attention. But this time all prompts were given in English ("The bird is above the plane," "The bird is in front of the plane," "The bird is on the plane," and "The bird is under the plane"). The negative feedback was given every time the child chose a wrong card. This study took a within-subject design, and all children experienced four types of feedback - Repetition, Negation, Description, and Demonstration (Table 1). There were 16 trials in total divided into four blocks, and each block tested all four spatial phrases (under, in front of, above, and on). The order of the trials was randomized within a block.

Figure 2. A scene from the Preposition Feedback Experiment. The child was asked to touch the image that corresponds to the English description given by the experimenter (e.g., The bird is on the plane).

3.1.4 Procedure

All participants met individually with the experimenter at their schools. Participants were seated in front of a 12.3-inch tablet (Microsoft Surface Pro 4) on which all visual stimuli were presented. The experimenter sat next to the child (Figure 2). The entire session took 10-20 minutes. Responses were coded online, but sessions with children were also videotaped in case further offline coding would be needed. Children were allowed to make up to three attempts on each trial. There were 16 trials (4 for each target word), and thus if a child had chosen the correct image on all trials, she would have only received positive feedback ("Yes! The bird is [on/in front of/above/under] the plane.") and would have given 16 responses only. If a child keeps choosing an incorrect image on all trials, on the other hand, she would be asked to respond 48 times in total.

3.1.5 Results

For each trial, participants' responses were coded in terms of how many times they chose a wrong answer. Figure 3 shows how the average number of errors for each type of negative feedback (Repetition, Negation, Description, or Demonstration) changed across the four blocks of the Test phase. The visual inspection of the data suggested that children chose wrong images more often when feedback was Repetition than when feedback was Negation, Description, or Demonstration, but none of the feedback reduced the number of errors. A two-way repeated measures ANOVA with Feedback Type (Repetition vs. Negation vs. Description vs. Demonstration) and Block (First vs. Second vs. Third vs. Fourth) as within-subject factors largely confirmed the pattern. The omnibus test found marginally significant effect of Feedback Type ($F(3,13) = 2.386$, $p = .084$). However, there was no significant effect of Block ($F(3,39) = .420$, $p = .739$), nor interaction between Feedback Type and Block ($F(9,117) = .245$, $p = .987$). Thus, Feedback Type might have affected the task performance, yet the number of errors did not decrease for any Feedback Type, and there was no indication that Feedback Type affected how well children learned English prepositions.

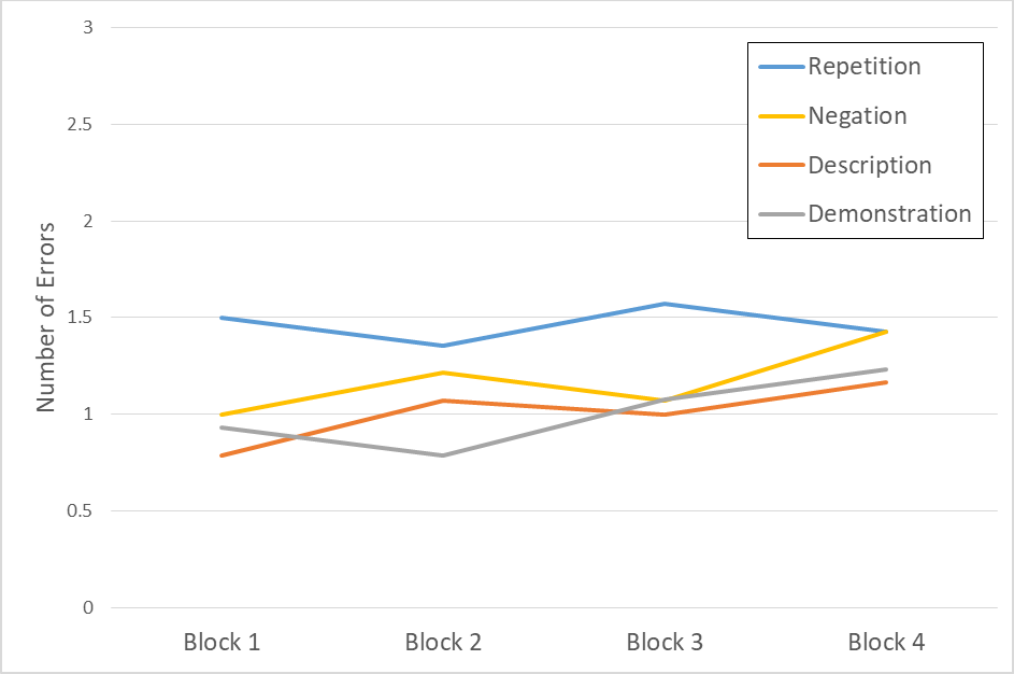

Figure 3. The number of errors across the four blocks.

3.1.6 Discussion

The Preposition Feedback Study tested how different types of negative feedback affected learning of spatial prepositions. Overall, the number of errors children made did not decrease across the four blocks, suggesting that none of the feedback given in the study helped children learn the words. We, however, found a trend in which children made more mistakes when negative feedback given was Repetition as opposed to Negation, Description, or Demonstration across all trials. One possible explanation to this pattern is Repetition being the least interesting among the four feedback types because the prompt was simply repeated. Perhaps children paid least attention to the task when the negative feedback was Repetition. Importantly, although the task performance was worst for words that were learned with Repetition, it was also very low for words that were learned with Negation, Description, and Demonstration. We suspected the task might have been too difficult and not engaging and thus not sensitive enough to detect the effects of feedback types not only because the performance was low but also because we have observed signs of fatigue and boredom (e.g., the child not looking at the tablet screen). Instead of continuing the data collection, we decided to conduct another less challenging study.

3.2 Verb Feedback Study

Based on what we learned from the Preposition Feedback Study, we designed the Verb Feedback Study that taught motion verbs to children. The Verb Feedback Study was similar to the Preposition Feedback Study, but we made a few changes in the design to ensure that the task was not too challenging for preschool children. First, in the Verb Feedback Study, children learned three words instead of four words. Second, we avoided teaching additional words by carefully choosing words and sentences to be taught. All target words were intransitive verbs (i.e., verbs that do not require objects in forming sentences). In the Preposition Feedback Study, children needed to learn not only the four prepositions but also the additional nouns “bird” and “plane” because spatial prepositions must describe relations between two objects. In contrast, intransitive verbs taught in the Verb Feedback Study (climbing, sliding, and falling) do not require objects of sentences. In addition, we used a common Turkish name, Elif, as subjects of sentences so that children did not need to learn English pronouns either. Third, we decreased the number of conditions by excluding Negation because the phrase used in Negation required children to understand the additional English word “not.” Thus, the Verb Feedback Study tested Repetition, Description, and Demonstration (Table 2). This study took a between-subject design in which children experienced only one type of negative feedback, which also made the task simpler and easier to follow.

3.2.1 Participants

Fifty preschoolers participated in Verb Feedback Study (*Age range* = 49-72 months; *Mean age* = 61.87 months; *SD* = 6.98; 22 females). Participants had no known vision or hearing impairments. Six additional children were also tested, but were excluded from analysis, as they did not finish the task.

3.2.2 Stimuli

The three target verbs taught in this experiment were sliding, climbing, and falling. Three video clips showing the motions corresponding to the target verbs were created to represent each of the words (Figure 4).

Table 2. Feedback provided by the experimenter in the Verb Feedback Study. All feedback phrases were given in English except for “Bir daha dene bakalım” which means “Give another try” in English.

	Feedback to correct response	Feedback to incorrect response
Repetition	Yes! Elif is sliding.	Hmm. Elif is sliding. *Bir daha dene bakalım.* Elif is sliding.
Description	Yes! Elif is sliding.	Hmm. Elif is [verb corrresponds to the image chosen by the child] (points to the image the child chose) *Bir daha dene bakalım.* Elif is sliding.
Demonstration	Yes! Elif is sliding.	Hmm. Elif is sliding (points to the correct image) *Bir daha dene bakalım.* Elif is sliding.

Figure 4. A screenshot from the Test phase of the Verb Feedback Study, showing the videos used to represent the three target verbs taught – falling (left), sliding (center), and climbing (right).

3.2.3 Design

There were four phases in the study: Learning, Practice, Test, and Transfer. Before the task started, the child was first presented with a short animation of a girl waving, and the experimenter introduced the girl as a friend named Elif.

In the Learning phase, the child watched Elif performing three target actions (sliding, climbing, and falling) one by one. In all animations, Elif was performing actions with the same background (yellow slide, orange sky, and grey ground). Each animation was played five times. During the presentation, the experimenter pointed to the animation and introduced the target verb by saying “*Bak, Elif böyle yapıyor*. Elif is sliding. *Bu harekete İngilizce* sliding *diyoruz*. Elif is sliding. *Şimdi benden sonra tekrar eder misin?* Sliding. Sliding. Elif is sliding.” (i.e., “*Look, Elif is doing this.* Elif is sliding. *In English, we call this action* sliding. Elif is sliding. *Now, can you repeat after me?* Sliding. Sliding. Elif is sliding.”). The order of the animations was randomized

across participants, and thus each child learned the verbs in one of the six orders (sliding-climbing-falling, sliding-falling-climbing, climbing-sliding-falling, climbing-falling-sliding, falling-sliding-climbing, and falling-climbing-sliding).

The Practice phase was included to ensure that the child knew how to complete the Test phase. Children were presented with two animations next to each other: one animation of a boy throwing a ball and another animation of a girl jumping. The experimenter asked "*Çocuk zıplıyor*" (i.e., "The child is jumping") in Turkish. The animations looped until the child responded.

In the Test phase, the same three animations used in the Learning phase were presented simultaneously (Figure 4). There were 12 trials in total divided into four blocks, and each block tested all three verbs (sliding, climbing, and falling). The order of the trials was randomized within a block. As in the Practice phase, the animations looped until the child chose one of the options. Importantly, however, in the Test Phase, when the child chose a wrong option on a particular trial, the child was given negative feedback and the same trial was repeated. Children were allowed to make up to three attempts on each trial. Hence, if a child chooses the correct image on all trials, she would not receive any negative feedback and make 12 responses only. If a child chooses an incorrect image on all trials, on the other hand, she would be asked to respond 36 times in total. There were three animations and thus there were six different ways in which they could be presented simultaneously on the tablet screen (sliding-climbing-falling, sliding-falling-climbing, climbing-sliding-falling, climbing-falling-sliding, falling-sliding-climbing, and falling-climbing-sliding). One of the six arrangements was randomly chosen for each time the child was prompted to choose an image.

The Transfer phase consisted of three trials on which the child was again asked to choose an animation that corresponds to each of the three verbs. This time, however, the options included novel animations. The first animation was the target in which an unfamiliar actor (a boy) performing the action referred by the verb with unfamiliar background (red slide, blue sky, and green ground). The second animation was a distractor in which a familiar actor (Elif) performing another action with unfamiliar background (red slide, blue sky, and green ground). The third animation was another distractor in which an unfamiliar actor (a boy) performing another action with familiar background. For example, when the child was asked to choose "sliding," she was presented with (1) an animation of a boy sliding on a red slide (target), (2) an animation of Elif climbing on a red slide (familiar actor distractor), and (3) an animation of a boy falling from a yellow slide (familiar background distractor). The order of three trials was counterbalanced across participants. The positions of the three animations were randomized across trials.

3.2.4 Procedure

As in the case of the Preposition Feedback Study, all participants met individually with the experimenter at their schools. Participants were seated in front of a laptop computer on which all visual stimuli were presented. The experimenter sat next to the child. Responses were coded online, but sessions with children were also videotaped in case further offline coding was needed. The entire session took 15-20 minutes.

3.2.5 Results

Participants' responses were coded in terms of how many times they chose a wrong answer. The number of errors was compared across the three conditions (Figure 5). Similar to Preposition Feedback Study, the Repetition condition yielded the highest

number of errors among the three conditions. However, a two-way repeated measures ANOVA with Feedback Type (Repetition vs. Description vs. Demonstration) as a between-subject factor and Block (First vs. Second vs. Third vs. Fourth) as a within-subject factor did not show significant results. The omnibus test found no significant effect of Feedback Type ($F(3,46) = .457$, $p = .636$), Block ($F(3,138) = .630$, $p = .597$), nor interaction between Feedback Type and Block ($F(6,138) = 1.229$, $p = .295$). Thus, Feedback Type might have affected the task performance, yet the number of errors did not decrease for any Feedback Type, and there was no indication that Feedback Type affected how well children learned English verbs.

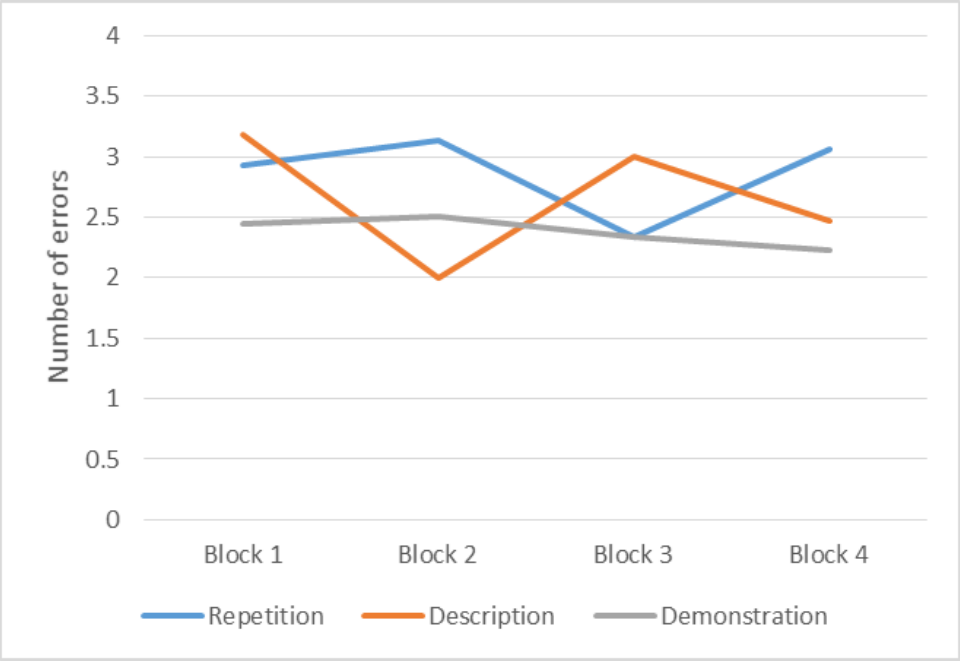

Figure 5. The number of errors across the four blocks of the Test phase.

The Transfer phase showed a clearer picture. All responses are coded in terms of the accuracy (1 = correct, 0 = incorrect). As shown in Figure 6, the number of correct responses was highest for children in the Demonstration condition followed by the Description condition, and then the Repetition condition. This pattern mirrors the number of errors made in the three conditions in the Test phase. To examine whether Feedback Type predicted the number of correct responses in the Transfer phase, we conducted another ANOVA including the number of questions that the child responded correctly on their first attempt in the very first questions asking for each of three verbs in the first block of the Test Phase (hereafter First Attempt) as a random factor and Age (in months) as a covariate. First Attempt was included as a random factor because the accuracy of the very first attempt reflects children's knowledge before they received any feedback, which we did not expect to differ across conditions. The omnibus test did not yield significant results, but there was a trend for Feedback Type affecting the performance in the Transfer Task ($F(2,14.23) = 2.707$, $p = .101$), suggesting that Feedback Type may predict how well children learn the target verbs in terms of whether they can generalize the meanings of words to novel contexts.

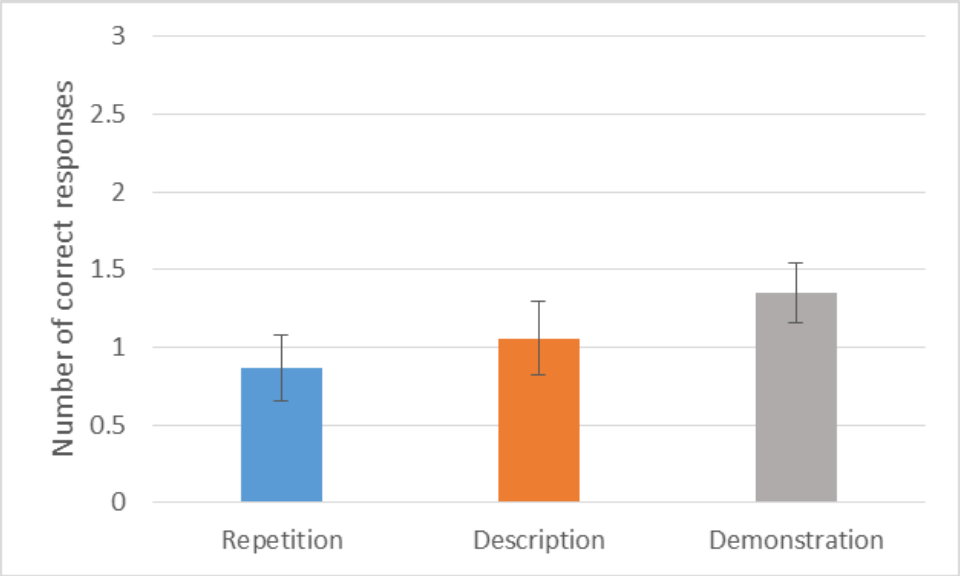

Figure 6. The number of correct responses in the Transfer phase.

Interestingly, however, our data also suggest the possibility that the advantage of the Demonstration feedback over the Repetition being different across participants. In fact, if we divide participants into the younger group and older group via median split (at 61 months of age; 4 children were excluded from this analysis as their exact ages were not available), an ANOVA including First Attempt as a random factor and Age (younger vs. older) as a covariate, yielded a three-way interaction predicting the accuracy in the Transfer phase ($F(2,27) = 3.34$; $p = .05$). Figures 7 and 8 show the performance of the younger and older children, respectively (Note that, only for the graphs, we divided children into low performers (0 or 1 correct responses) and high performers (2 or 3 correct responses) based on their accuracy on the first questions, i.e., First Attempt; the statistical analysis reported here did not regroup children that way and First Attempt was kept as a continuous variable).

Among younger children (Figure 7), the Repetition was more effective for those who performed poorly on the first questions in the Test phase and thus received negative feedback. On the other hand, the Description appeared to be more beneficial for those who already know the answer on the first questions in the Test phase. On average, the same was true for the Demonstration though the difference based on the accuracy on the first questions was minimal for this condition.

As shown in Figure 8, the pattern was flipped for the older children. Among the older children, the Repetition appeared to be more beneficial for those who performed well on the first questions whereas the Description was more beneficial for those who performed poorly. On average, the Demonstration was more beneficial for the older children who performed well before receiving any negative feedback, but as in the case of the younger group, the difference based on the accuracy on the first questions was minimum among the three conditions.

In summary, the results suggest that, with regard to the Transfer phase, the Demonstration may be effective regardless of children's age or their initial performance whereas the Repetition and Description seem to affect children differently. The pattern, however, must be interpreted with caution as the number of participants is low.

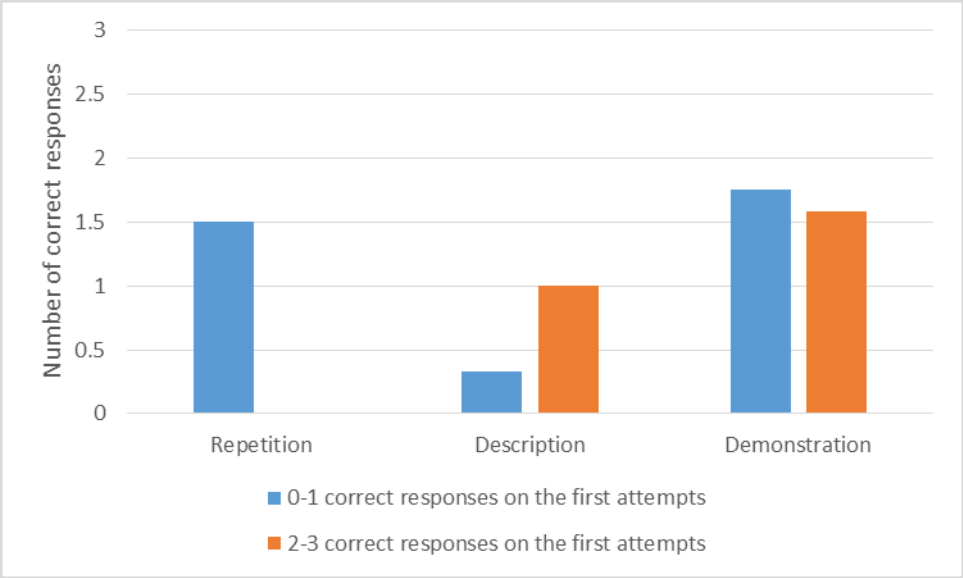

Figure 7. The number of correct responses in the Transfer phase for children who were 61 months old or younger.

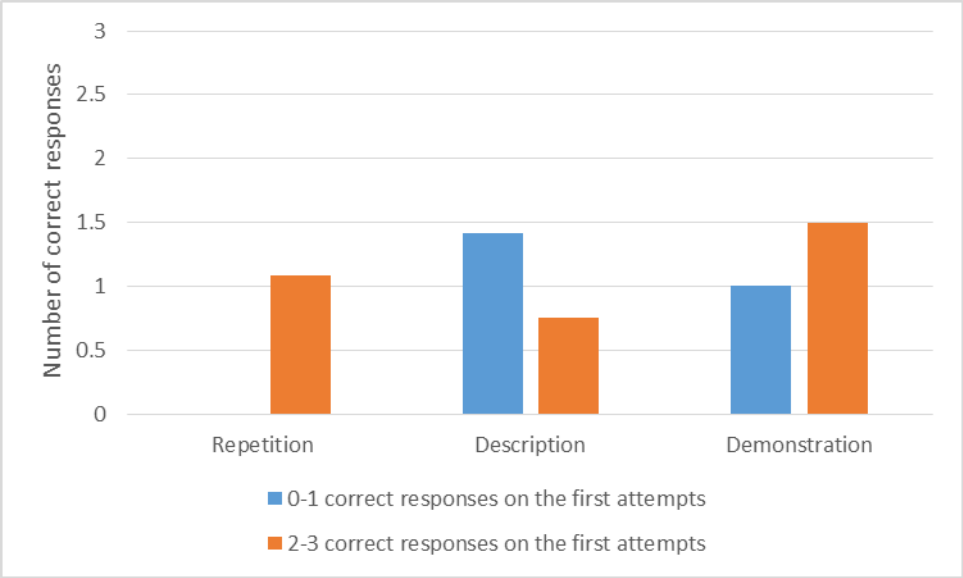

Figure 8. The number of correct responses in the Transfer phase for children who were older than 61 months of age.

3.2.6 Discussion

The Verb Feedback study tested whether types of negative feedback predict how well children learn motion verbs. We predicted that the Description and Demonstration conditions would entail more learning than the Repetition condition, since they provide children with information they can use in the upcoming trials in contrast to mere repetition of the prompt. During the Test phase in which the negative feedback was given, we did not observe significant increase in children's performance. However, our results from the Transfer phase suggest marginal effect of feedback types on verb learning, driven potentially by the Demonstration condition. Further, in concert with the pattern found in the Preposition Feedback study, the Repetition condition resulted in the lowest level of learning. However, these patterns must be interpreted with caution as we

found no statistically significant difference between these three types of feedback. A follow-up analysis that divided children into younger and older groups revealed a three-way interaction among the feedback type (Repetition vs. Description vs. Demonstration), age (younger vs. older), and the accuracy on the first questions in the Test phase. The Demonstration appeared to be most effective regardless of children's age or their initial performance. The Repetition and Description, on the other hand, seem to affect children differently. As our study aimed to serve as a pilot for a robot version of the study, we tested a relatively small number of participants. The possibilities discussed above must be evaluated with a larger number of participants.

In addition to the relatively small sample size, the lack of significant results in the Test phase may be due to our study design. Most research conducted with adults revealed the advantage of explicit verbal feedback over implicit feedback in second language learning (e.g., Ellis, Loewen, & Erlam, 2006). Our study followed the patterns observed in L2 classroom and avoided highly explicit negative feedback. For example, the experimenter did not say "No" but made a sound of disapproval. Therefore, feedback we tested might have been too implicit. It is important to emphasize that, among the three, the Demonstration condition was the most explicit as the child was immediately able to see and hear the correct response, compared to the Description and Repetition conditions. This may have resulted in the slight yet insignificant advantage of the Demonstration condition on learning. Our participants, they might have experienced difficulty in understanding whether their response is correct or not, prior to listening to the rest of the feedback.

Further, most research on the effects of feedback on children's second language learning has been focusing on elementary school-aged children and older (Lyster & Saito, 2010), whereas our sample consists of preschoolers. Young age of children, coupled with the repetitive nature of the task might have led to low engagement levels in children which in turn led to lower levels of learning across conditions. In fact, our analysis on the accuracy in the Transfer phase suggests that the effectiveness of the feedback – repetition, description, and demonstration – may differ across age groups. Previous research mostly studied the effects of feedback in more interactional contexts where children received feedback in mutual dialogues with an adult over the course of a few weeks (e.g., Mackey & Oliver, 2002). To isolate the individual effects of each feedback type in a fully experimental setting, we followed a strict protocol where the experimenter provides one-way feedback without engaging in a dialogue with children. This can be seen as a challenge for young children as it vastly differs from their natural social learning settings. Although no direct measures were taken, we observed that children had difficulties in focusing on the task and kept seeking for the experimenter's comments on their performance. This often led the experimenter to motivate children to continue, which may have also interfered with the feedback she was providing.

As mentioned earlier, a NAO robot version of this study has been designed and will be conducted in the coming months. The study with a NAO would not only inform us what feedback can be used in lessons led by a robot tutor, but would give us important insights on influence of feedback types on L2 learning because robots allow us to conduct ecologically valid yet systematic experiments. While an unnatural interaction with an adult can feel artificial for children and potentially hinders their learning, robots can provide a strict experimental manipulation while drawing children's attention on the learning material in a novel social setting.

4 Gesture Study

As in the case of feedback, there are different types of gestures. Previous research found that young children are sensitive to gestures. For example, children show sensitivity to an adult's pointing gesture and shift their attention in the direction of pointing (Rohlfing, Longo, & Bertenthal, 2012). There is also evidence that iconic gestures can support L2 vocabulary acquisition in children (Tellier, 2008; Rowe, Silverman, & Mullan, 2013). However, it is still unclear (1) when and how gestures facilitate L2 word learning and (2) which types of gestures are more effective in teaching words. To answer these questions, we examined children's word learning with a human experimenter in three different conditions: deictic gesture condition, iconic gesture condition, and no gesture condition. We also aim to test effects of beat gestures in future though the condition was not included in the current study because determining effects of deictic and iconic gestures is most critical for the project, and the comparison between the two gesture conditions and no gesture condition is sufficient to evaluate whether body movements in general enhances word learning. In this study, a human experimenter taught four pairs of adjectives (small and big, wide and narrow, high and low, and tall and short). These adjectives were chosen because they are easy to gesture and those gestures can be easily performed both by a human experimenter and NAO.

4.1 Participants

Thirty four 5-year-olds participated in the Gesture Study (*Mean age* = 65.64 months; *SD* = 4.84; 15 females). Among them, 14 children participated in the Deictic Gesture condition (*Mean age* = 64.63 months; *SD* = 5.37; 5 females) and 20 participated in the Iconic Gesture condition (*Mean age* = 68.33 months; *SD* = 1.15; 9 females). Participants had no known vision or hearing impairments.

4.2 Stimuli

Four pairs of adjectives – small and big, wide and narrow, high and low, and tall and short – were taught in this experiment. From the measurement words (i.e., measurable attitudes) in the curricula for kindergarten math in the Common Core in the US, we first selected six pairs of adjectives (i.e., 12 adjectives) for which generating iconic gestures seemed fairly easy. To ensure that all adjectives can be represented well with iconic gestures that can be performed by NAO, we conducted an online survey with a separate group of 25 adults (*Mean age* = 33.19 years; *SD* = 6.50; 10 females). On a 5-point scale, participants rated how well the gesture that appears in a video clip represents a specific adjective (e.g., high). Based on the data, the four pairs of adjectives (small and big, wide and narrow, high and low, and tall and short) were chosen as target words for the main experiment. Gesture-adjective pairs were on average rated as 3.3 corresponding to moderately well. We additionally made sure that these adjectives are fairly balanced in terms of word frequency, concreteness, familiarity, and imagibility (Table 3). An image of common objects was created to represent each pair of adjectives. These images were an image of two balls (big and small), two doors (wide and narrow), two kites (high and low), and two flowers (tall and short; see Figure 9).

Table 3. Frequency, concreteness, familiarity, and imagibility of the adjectives used in the Gesture Study. The words are ranked according to Educator's Word Frequency Guide (Zeno, Ivens, Millard, & Duvvuri, 1995) and MRC Psycholinguistic Database.

	WFG Database		MRC Psycholinguistic Database			
	All	1st-6th	concreteness	familiarity	imagibility	frequency
narrow	988	362	372	546	491	63
Wide	1633	691	348	569	455	125
Low	2103	844	322	580	378	174
Tall	1899	1165	439	585	514	55
Short	2795	1070	351	586	431	212
High	5863	2380	371	612	463	497
Big	8973	7065	-	640	463	360
Small	9561	4031	402	616	447	542

Table 4. The list of adjectives taught in the Gesture Study, together with the objects used in describing the meanings of the adjectives and iconic gestures used in the Iconic Gesture condition.

	Adjective	Object	Iconic gesture
1a	big	ball	Horizontally extending the arms to the sides, parallel to the ground
1b	small	ball	Holding the hands in a sphere-like shape in front of the chest
2a	tall	flower	Moving the right arm up above the head level, the palm facing the ground
2b	short	flower	Moving the right arm down to the side of the body, the palm facing the ground
3a	high	kite	Moving the right arm up above the head level, the palm facing the front
3b	low	kite	Moving the right arm down, palm facing down, fingers pointing the ground
4a	wide	door	Holding the hands in front of the chest, both perpendicular to the ground, far apart from each other
4b	narrow	door	Holding the hands in front of the chest, both perpendicular to the ground, close to each other

4.3 Design

Our primary concern was to compare deictic and iconic gestures, but we also included the Highlight condition to examine whether gestures are at all effective in facilitating word learning. We used a mixed design in which children went through two conditions: one of the gesture conditions (Deictic or Iconic) and the Highlight condition. No gesture was performed in the Highlight condition, and a red square appeared around the object to draw children's attention to the image (Figure 9). Therefore, this study included two types of gestures (Deictic or Iconic) and two ways of directing children's attention to the referents of objects (Gesture vs. Highlight). Inclusion of the Highlight condition allowed us to understand whether findings in the Deictic and Iconic Gesture conditions are due to unique influence of bodily gestures or the results of drawing attention to the referents of words in general.

There were three blocks per condition. In each condition, children learned two pairs of adjectives, i.e., four adjectives. Thus, children learned four pairs of adjectives, i.e., eight adjectives, throughout the experiment. The adjective pairs were counterbalanced. Half of the children learned two pairs of adjectives (e.g. big-small, high-low) in the Gesture condition, whereas half of them learned the same pairs in the Highlight condition. In addition, children were taught names of objects (e.g., ball) that were used to teach adjectives.

There were three phases within each of the three blocks: Noun Learning, Adjective Learning, and Test. In the Noun Learning phase, children were taught the name of an object (e.g., ball) to be used to represent the pairs of adjectives (e.g., small and big). Each pair of adjectives were taught with a single object (see Figure 9). Then, in the Adjective Learning phase, the objects were presented one by one. In the Deictic Gesture condition, the experimenter pointed to the object on the tablet screen while teaching the adjective. In the Iconic Gesture condition, the experimenter performed an iconic gesture (Table 4) while teaching the adjective. In the Highlight condition, the experimenter taught the adjective without any bodily gesture, but a red square appeared around the object to draw children's attention to the image. In the Test phase, the child was once again presented with the image of two objects (e.g., a small ball and a big ball), and was asked to point to the object that corresponds to the learned adjective (e.g., small and big).

After completing three blocks for the Deictic/Iconic Gesture condition and three blocks for the Highlight condition, the child was asked to complete a transfer task. In this transfer task, the child was presented with a series of new images representing the same set of eight adjectives but with different objects. The child was again asked to point to the object that corresponded to each of the eight adjectives. The transfer task was included as a stringent test to evaluate whether children really learned the words.

4.4 Procedure

As in the case of the feedback studies, all participants met individually with the experimenter at their schools. The child was seated in front of a 13-inch screen on which all visual stimuli were presented, and the experimenter sat across from her. The entire session took 15-20 minutes. Responses were coded online, but sessions with children were also videotaped in case further offline coding was needed.

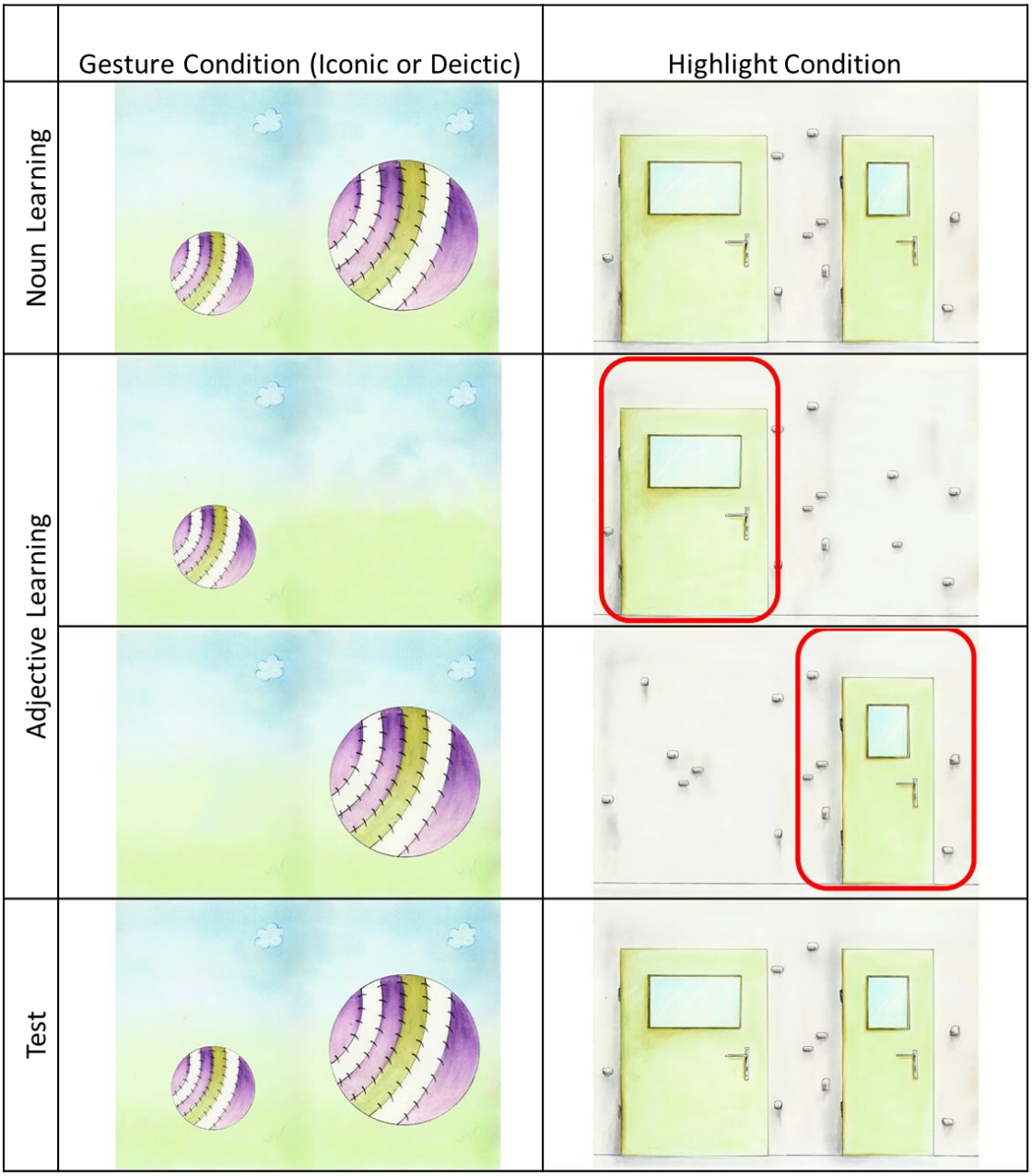

Figure 9. Examples of the images appeared on the tablet in the Gesture Study. In all conditions (Iconic Gesture, Deictic Gesture, and Highlight), the child was first shown an image showing two different versions of the same object (e.g., ball, door), and learned the English noun for the object. Then, the objects were presented one by one to introduce the target adjectives (e.g., small, big, wide, narrow). In the two gesture conditions (the left column), the experimenter performed gestures while introducing the adjectives. In the Highlight condition (the right column), no gesture was performed and red rectangles appeared around the object.

4.5 Results

For each trial, participants' pointing responses were coded in terms of whether they pointed to the correct answer or not (1 = correct, 0 = incorrect). The distribution of scores approximately followed the normal distribution within each condition, and thus the data were analyzed using an ANOVA. Two separate two-way mixed ANOVAs with Gesture Type as a between-subject variable (Iconic vs. Deictic) and Modality as a

within-subjects variable (Gesture vs. Highlight) was used to analyze the responses – one on children's responses during the main task, and the other on children's responses for the transfer task.

First, ANOVA on children's responses during the main task revealed a significant effect of Modality ($F(1,33) = 4.11$, $p = .05$), where children performed better on the Highlight condition ($M = .73$, $SD = .25$), compared to the Gesture condition ($M = .66$, $SD = .19$). Further, there was a trending interaction between Gesture and Modality ($F(1,33) = 2.69$, $p = .11$). The interaction was driven by the fact that children performed better on the Highlight condition when the Highlight condition followed the Deictic Gesture condition ($M = .83$, $SD = .23$), compared to when it followed the Iconic Gesture condition ($M = .69$, $SD = .25$; see Figure 10). There was no statistically significant effect of Gesture Type ($F(1,33) = 1.04$, $p = .32$). A mixed ANOVA on the transfer task did not reveal any significant effect of Gesture Type ($F(1,32) = .54$, $p = .46$), Modality ($F(1,32) = 1.52$, $p = .23$), nor their interaction ($F(1,32) = .39$, $p = .53$).

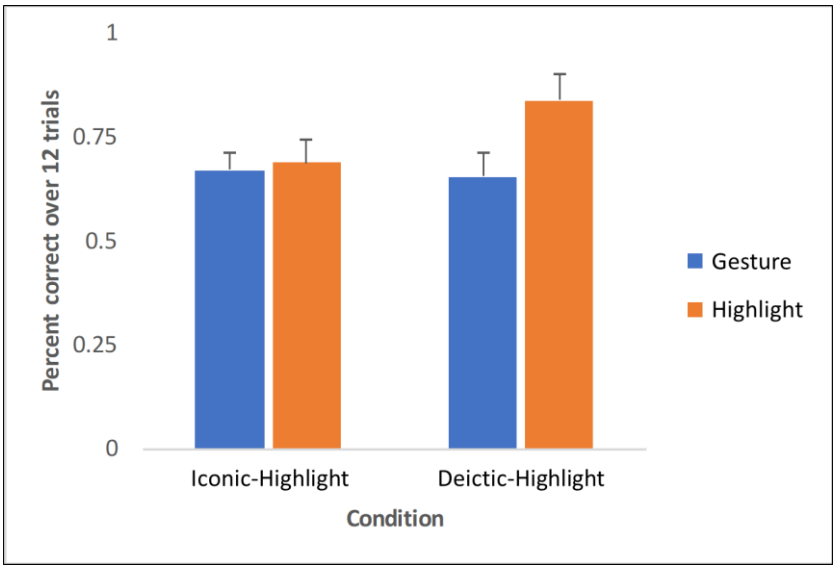

Figure 10. The percentage of correct responses during the main task. Error bars indicate the standard error.

4.6 Discussion

Gesture Study tested whether gestures facilitate L2 word learning and whether a specific type of gesture, deictic or iconic, was more effective than the other. Contrary to some of the prior literature, we did not find significant differences between the gesture conditions. It is important to note that prior work primarily focused on learning novel nouns, whereas here we focus on learning of spatial adjectives (McGregor, 2008). Further, we also found that children performed better when presented with attention highlighters compared to gestures. Recent work suggests that the role of gesture might vary with prior knowledge. For children with low prior knowledge, gesture might not be as effective as concrete actions performed with real objects (Congdon, 2016). The current study tested children with minimal knowledge of English. Based on teacher reports, children did not know any of the adjectives that were taught to them. Similarly, for these children who are being exposed to words in L2 for the first time, gestures might be too abstract to aid learning. These children might instead benefit more from attention highlighters that are similar to concrete actions in that they directly attract children's visual attention to the object mentioned in speech. Relatedly, our findings

also move the prior literature one step forward by suggesting that gestures' role might not be the same as other attention grabbers. Finally, prior work examining the role of gestures in verb learning highlighted the importance of children doing the gestures as opposed to observing them (Wakefield, Hall, James, & Goldin-Meadow, 2018). Future studies should manipulate whether this distinction plays an important role in learning of spatial adjectives as well.

Our results suggest that a combination of deictic gestures followed by attention highlights might bring the most desirable learning outcomes for children. This might be because both the Deictic Gesture condition and the Highlight condition attract children's attention to the screen, deeming it easier to associate the adjectives with their corresponding visuals. In the current design, the Highlight condition always followed one of the gesture conditions. Thus, we cannot clearly state whether the Highlight condition on its own would bring desirable results as well or whether this condition followed by deictic gestures led to the best outcomes. Currently, we are examining this possibility where we present the Highlight condition before the gesture conditions.

Finally, it should also be noted that there was a great variability in children's performance. Our ongoing analyses examine whether the variability in children's responses can be explained by their age and other individual differences measures, such as their working memory capacity or attentional focus.

5 Discussion and Conclusion

This deliverable reported three experimental studies that examined whether and how types of feedback and of gestures affect L2 word learning in preschoolers. In the first feedback study which taught English prepositions to Turkish-speaking children, we did not find specific types of negative feedback to improve the learning outcome though simply repeating the prompt, which was observed in L2 classroom, appeared to be the least favorable strategy. The second feedback study with motion verbs found similar patterns. In addition, the results seem to suggest that demonstrating the correct answer may be the most effective when providing negative corrective feedback to young children. The third study tested effects of gesture types in L2 word learning. In this study, types of gestures, deictic and iconic, did not predict how well children learned novel words. To our surprise, we found that directing attention to the referent of a word with a red square (Highlight condition), especially when it followed after Deictic Gesture condition, to be the most effective method. It is important to note that these findings may be due to the tested word type (i.e., adjectives) or the target population (i.e., children with no prior knowledge of English), and thus the generalizability of the results should be further examined in future.

The human experiments presented in this deliverable provide important insights for developing lessons within robot-assisted L2 learning. As noted earlier, we plan to conduct robot versions of Verb Feedback Study and Gesture Study in the coming months to examine whether the findings reported here apply for lessons that are led by a NAO. These future studies will allow us to evaluate whether types of feedback and of gestures affect L2 word learning differently when the lessons are provided by a human or by a robot.

References

Carroll, S., & Swain, M. (1993). Explicit and implicit negative feedback: An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition, 15*(3), 357-386.

Congton, E. L. (2016). *Learning mathematics through action and gesture: Children's prior knowledge matters* (Doctoral Dissertation, The University of Chicago).

Conti, D., Di Nuovo, A., Cirasa, C., & Di Nuovo, S. (2017). A comparison of kindergarten storytelling by human and humanoid robot with different social behavior. *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human–Robot Interaction* (pp. 97-98), March 6-9, 2017, Vienna, Austria.

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, *28*(2), 339-368.

Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., … Breazeal, C. (2016). Affective personalization of a social robot tutor for children's second language skills. *Proceedings of the 30th AAAI Conference on Artificial Intelligence* (pp. 3951-3957), February 12-17, 2016, Phoenix, AZ.

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological bulletin, 137*(2), 297.

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

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching*, *46*(1), 1-40.

Macedonia, M., Müller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human brain mapping, 32*(6), 982-998.

Mackey, A., & Oliver, R. (2002). Interactional feedback and children’s L2 development. *System, 30*(4), 459-477.

McGregor, K. K. (2008). Gesture supports children's word learning. *International Journal of Speech-Language Pathology*, *10*(3), 112-117.

Rohlfing, K. J., Longo, M. R., & Bertenthal, B. I. (2012). Dynamic pointing triggers shifts of visual attention in young infants. *Developmental Science, 15*, 426-435.

Rowe, M. L., Silverman, R. D., & Mullan, B. E. (2013). The role of pictures and gestures as nonverbal aids in preschoolers’ word learning in a novel language. *Contemporary Educational Psychology, 38*(2), 109-117.

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*, 661-699.

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, *8*(2), 219-235.

Valenzeno, L., Alibali, M. W., & Klatzky, R. (2003). Teachers’ gestures facilitate students’ learning: A lesson in symmetry. *Contemporary Educational Psychology, 28*, 187-204.

Wakefield, E. M., Hall, C., James, K. H., & Goldin-Meadow, S. (2018). Gesture for generalization: gesture facilitates flexible learning of words for actions on objects. *Developmental Science*.

Zeno, S., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Touchstone Applied Science Associates.