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

D7.4 Evaluation Report Storytelling Domain

Due Date: **31/12/2018**
Submission Date: **31/12/2018**

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

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

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

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

Contents

Executive Summary

In the revised objectives of L2TOR (see the mid period review report) we set out the plan to conduct a series of small-scale studies to further explore several questions that emerged during the study. Answers to these questions can further advance the development of a language tutoring robot platform. This deliverable is used to report the results of these smaller scale studies conducted by the partners in the consortium. In the introduction we specify for each small-scale study listed in the revised objectives where the results are reported (in the current deliverable or in a different deliverable).

The studies included in this deliverable span a range of topics. Two studies (sections 2 and 5) focus on whether and how different types of gestures (iconic, deictic, and beat) and the match between words and iconic gestures (i.e., how well iconic gestures represent the meanings of corresponding words) affect word learning. The studies found no effect of gesture types, but the match between words and iconic gestures predicted whether children learned the words or not.

In section 3 we describe detailed results of a study where the robot was used to teach Turkish-Dutch children Dutch vocabulary. Results show that, using the current setting, children learned better with a robot speaking only Dutch than with a robot that provided Turkish translations of the word. We discuss the meaning of the findings and the limitations of the current design. Another study focused on feedback during learning with the robot. We tested the hypotheses that children will learn more when the robot provides feedback and that they will learn more when the robot provides the type of feedback that human teachers prefer using.

In section 4 we describe the design and rationale for this study that has recently been conducted. We plan to analyse the data of this study in January 2019.

In section 6 we report about a review study looking into the effect of a robot platform on learning. Results show that there is some evidence that social and agentic nature of robots promotes responses that are conducive to learning. Moreover, for young children shorter robots are usually preferred, as it is assumed that younger children will feel more comfortable with them.

Finally, in section 7 we report on an additional small-scale study that we conducted which was not included in the revised objectives. In this study, we compared whether children attribute similar importance to non-verbal cues (i.e., eye gaze and pointing) of a robot as to those of a human. We found that children do not differ in their reliance on non-verbal cues of a robot versus those of a human, and that differences in anthropomorphism interacted with children's reliance on a robot's pointing behaviours.

Principal Contributors

UU: Ora Oudgenoeg-Paz, Hanneke Leeuwestein, Paul Leseman, Josje Verhagen, Rianne van den Berghe
KOC: Junko Kanero, Cansu Oranç, Özlem Ece Demir-Lira, Sümeyye Koşkulu, Tilbe Göksun, Aylin C. Küntay
TIU: Mirjam de Haas, Emiel Krahmer, Paul Vogt, Bram Willemsen, Jan de Wit
PLYM: Tony Belpaeme, James Kennedy, Chris Wallbridge

Revision History

Version 1.0 (RB 31-12-2018)
This is the first version.

1 Introduction

The title of this deliverable is Evaluation report storytelling domain. However, given the changes we applied to the planning of the project (see also the mid period review report) we changed the content of this document. The storytelling domain is no longer included in the project. Therefore, this deliverable is used to report the results of smaller scale studies conducted by the partners in the consortium. The reviewers recommended that we devote efforts in our final 18 months to conducting a reduced large-scale study and, in addition, a few smaller scale projects were planned to address several issues we encountered while designing the L2TOR system. The results of several of these studies were reported in other deliverables. The remaining studies are included in this deliverable. Below is the list of small studies provided and approved in the revised objectives. For each study we note in what deliverable the results are reported.

1. Gesture type study: how do human tutors gesture? - results are reported in deliverable 1.3. See section 2 of this deliverable for an update on the results.
2. Robot tutoring of Dutch in Turkish migrant communities – initial findings were reported in D7.1, In section 3 of this deliverable we report our final analyses of these data.
3. The impact of a robot's physical limitations on gesture comprehensibility – results are reported in deliverable 6.3
4. Teacher feedback during L2 learning – this study has been conducted and the data will be analysed in the beginning of January. In section 4 of this deliverable we describe the method we used and research question addressed.
5. The impact of affect detection and adaptation on learning – results are reported in deliverable 5.3
6. The impact of short-term and long-term adaptation on learning – this study has not been conducted as the planning of WP 5 has been adjusted. These adjustments are discussed in detail in deliverable 5.3
7. Verbalisation of system knowledge – Results are presented in deliverable 5.3
8. Gesture production, rating and learning study – results are reported in section 5 of this deliverable
9. Encouraging the Production of Spatial Concepts in L2 – results are reported in deliverable 6.3.
10. Influence of robot platform on learning – results are reported in section 6 of this deliverable.

2 Gesture type study: how do human tutors gesture? – an update of the results

We have added more participants to the gesture study reported in Deliverable 1.3 and completed data collection. In this study, we conducted three experiments to investigate the effects of different gesture types on children's L2 word learning. In Studies 1 and 2, as proposed in the revised objectives, we tested different types of gestures – iconic gestures (Studies 1 and 2), deictic gestures (Study 1), and beat gestures (Study 2) – performed by a human tutor. We also conducted an additional study with the NAO to see if the results of Study 1 were replicated when a robot played the role of the tutor (Study 3). In the last section of this report, we compare Studies 1 and 3 and discuss differences between the human and robot versions of the study.

As described in Deliverable 1.3, children learned four pairs of English measurement words (e.g., small and big) with images of objects representing these words (e.g., small ball and big ball) presented on a laptop screen. All children experienced two learning conditions – one of the two Gesture conditions (Iconic or Deictic; Figures 1A and 1B) + the On-Screen Highlighter condition (Figure 1D; called the "Highlight condition" in Deliverable 1.3) in Studies 1 and 3, and the Iconic Gesture condition + the Beat Gesture condition in Study 2 (Figure 1C). In the Iconic condition, the tutor produced an iconic gesture representing the measurement word (Figure 1A). In the Deictic condition, the tutor whole-hand pointed to the corresponding object on the screen (e.g., small ball; Figure 1B). In the Beat condition, the tutor made rhythmic hand movements (moving her hands up and down as she spoke; Figure 1C). During the test, children were asked to point to the picture corresponding to the target measurement word.

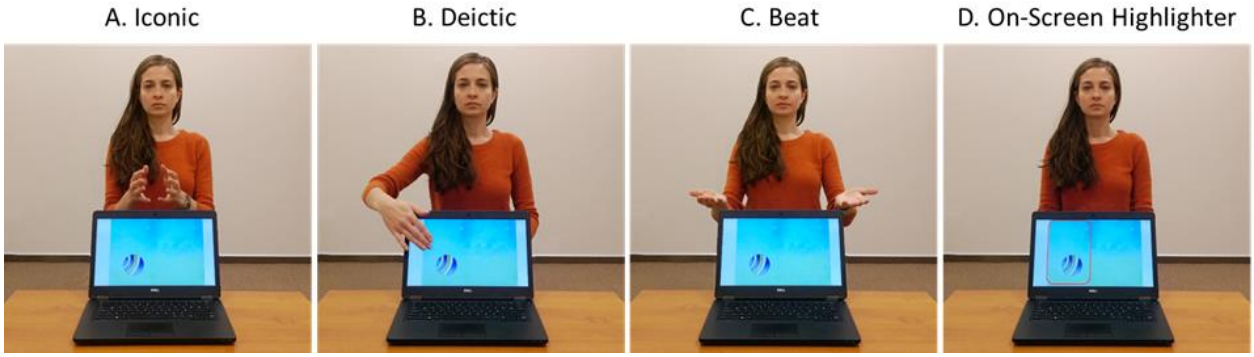

Figure 1. Examples of the gestures performed by the tutor and the images presented on the laptop screen. In the Iconic, Deictic, and Beat Gesture conditions, the tutor performed a gesture while the corresponding object (e.g., small ball) was presented on the screen. In the On-Screen Highlighter condition, the same image was presented but a red rectangle appeared to highlight the object, and the tutor did not move.

The final sample consists of 100 Turkish-speaking preschoolers: 41 in Study 1, 22 in Study 2, and 37 in Study 3. For all studies, we ran generalized linear mixed effects model (GLMM) analyses because our dependent variable was binary (0 = Incorrect, 1 = Correct). We used the glmer function in *lme4* package on R (Bates, Maechler, Bolker, &

Walker, 2014). We used the *lmerTest* package to obtain p values for the fixed effects (Kuznetsova, Brockhoff, & Christensen, 2017). In all models, Subject and Word were included as random effects, with random intercepts allowed. In Studies 1 and 3, Condition had three levels (i.e., Iconic gesture, Deictic gesture, On-Screen Highlighter), and thus we built two models to test all possible pairs of Conditions (Iconic vs. Deictic, Iconic vs. On-Screen Highlighter, and Deictic vs. On-Screen Highlighter), one with Iconic as the reference group and the other with On-Screen Highlighter as the reference group.

Study 1: Iconic and Deictic Gestures by a Human Tutor

Study 1 tested whether iconic and deictic gestures performed by a human tutor differently affect children's word learning. The final sample consists of 41 preschoolers (*Age range* = 57.36-77.04 months; *M* = 67.60 months, *SD* = 4.99).

As shown in Table 1, learning outcomes in the two Gesture conditions were not significantly different from each other (*Mean Score*$_{Iconic}$ = 8.43; *SD* = 2.43; *Mean Score*$_{Deictic}$ = 7.89; *SD* = 2.22), whereas children performed better in the On-Screen Highlighter condition (*Mean Score*$_{Highlighter}$ = 9.07; *SD* = 2.94) than in the Deictic condition (see Table 2). We also compared the On-Screen Highlighter condition with the two Gesture conditions combined. Confirming previously reported results (D1.3), children performed better in the On-Screen Highlighter condition than the two Gesture conditions combined (*Mean Score*$_{Gesture}$ = 8.20; *SD* = 2.33; see Table 3). In both analyses, children gave more correct responses in Block 3 compared to Block 1, indicating learning throughout the experiment.

These results indicate that, when performed by a human tutor, types of gestures did not make a difference in children's L2 word learning. Furthermore, children gave more correct responses when their attention was drawn to the laptop screen where learning material was presented (i.e., On-Screen Highlighter condition), than when the human tutor used gestures.

Table 1
Study 1: Fixed and random effects for the GLMM predicting children's responses in the human-led lesson (Reference group: Iconic) (N=41)

Fixed effects	**Estimate**	**SE**	**Wald Z**	***p***
Intercept	0.96	0.29	3.35*	< .001
Highlighter (vs. Iconic)	0.27	0.20	1.33	.18
Deictic (vs. Iconic)	-0.39	0.29	-1.37	.17
Block 2 (vs. Block 1)	0.24	0.19	1.30	.19
Block 3 (vs. Block 1)	0.53	0.19	2.71*	.01
Random effects	**Variance**	**SD**		
Subject	1.15	1.07		
Word	0.10	0.32		

* $p < .05$

Table 2
Study 1: Fixed and random effects for the GLMM predicting children's responses in the human-led lesson (Reference group: Highlighter) (N=41)

Fixed effects	**Estimate**	**SE**	**Wald Z**	***p***
Intercept	1.23	0.26	4.74*	< .001
Deictic (vs. Highlighter)	-0.66	0.23	-2.92*	.003
Iconic (vs. Highlighter)	-0.27	0.20	-1.33	.18
Block 2 (vs. Block 1)	0.24	0.19	1.30	.19
Block 3 (vs. Block 1)	0.53	0.19	2.71*	.01
Random effects	**Variance**	**SD**		
Subject	1.15	1.07		
Word	0.10	0.32		

* $p < .05$

Table 3
Study 1: Fixed and random effects for the GLMM predicting children's responses in the human-led lesson (Highlighter vs. Gesture conditions combined) (N=41)

Fixed effects	**Estimate**	**SE**	**Wald Z**	***p***
Intercept	1.23	0.26	4.75*	< .001
Gesture (vs. Highlighter)	-0.44	0.16	-2.81*	.005
Block 2 (vs. Block 1)	0.24	0.19	1.30	.19
Block 3 (vs. Block 1)	0.52	0.19	2.7*	.01
Random effects	**Variance**	**SD**		
Subject	1.14	1.07		
Word	0.1	0.32		

* $p < .05$

Study 2: Iconic and Beat Gestures by a Human Tutor

According to the results of Study 1, the Iconic and Deictic conditions did not differ in terms of children's learning outcomes although the Deictic condition yielded significantly lower accuracy than the On-Screen Highlighter condition. To further understand the effects of iconic gestures, we conducted Study 2 in which iconic gestures were compared with beat gestures in a within-subjects design where the two conditions were counterbalanced across participants. Thus, a new group of preschoolers learned the same set of measurement words from a human tutor who used iconic and beat gestures. In total, 22 children participated (*Age range* = 59.33-81.87 months; *M* = 71.51 months, *SD* = 6.71).

There was no statistically significant difference between the two conditions (see Table 4; $Mean\ Score_{Iconic}$ = 8.27; *SD* = 3.07; $Mean\ Score_{Beat}$ = 7.55; *SD* = 3.39). No effect of condition order (hereafter Order; Iconic-Beat vs. Beat-Iconic) nor difference across blocks (Block 1 vs. Block 2 vs. Block 3) was observed. In sum, iconic and beat gestures led to comparable learning outcomes in the human-led L2 word lessons for children.

Table 4
Study 2: Fixed and random effects for the GLMM predicting children's responses in the human-led lesson (Iconic vs. Beat) (N=22)

Fixed effects	**Estimate**	**SE**	**Wald Z**	***p***
Intercept	0.61	0.45	1.35	.18
Iconic (vs. Beat)	0.27	0.21	1.30	.19
Order: Iconic-Beat (vs. Beat-Iconic)	0.15	0.44	0.34	.74
Block 2 (vs. Block 1)	-0.13	0.25	-0.52	.61
Block 3 (vs. Block 1)	0.13	0.25	0.51	.61
Random effects	**Variance**	**SD**		
Subject	0.81	0.90		
Word	0.53	0.73		

* $p < 0.05$

Study 3: Iconic and Deictic Gestures by the Robot Tutor

Because Study 1 elicited some significant results across conditions, we decided to replicate the study with the robot tutor, with 37 children (*Age range* = 56.64-77.90 months; *M* = 69.86 months, *SD* = 4.18). A flaw of Study 1 was that On-Screen Highlighter always came as the second condition, and thus in Study 3, we counterbalanced the order between the Gesture (i.e., Iconic or Deictic) and On-Screen Highlighter conditions as we did in Study 2.

In concert with the human version of the study (Study 1), the Iconic and Deictic Gesture conditions were not significantly different from each other (Table 5). However, in contrast to Study 1 where we found the On-Screen Highlighter condition to be significantly better than the Deictic condition, in Study 3, we found that the On-Screen Highlighter condition (*Mean Score*$_{Highlighter}$ = 9.95; *SD* = 2.50), was significantly better than the Iconic condition (*Mean Score*$_{Iconic}$ = 9.00; *SD* = 2.42; see Table 6).

Order also had a significant effect, indicating that children performed better in the condition that was presented as the second condition. Children also gave significantly more correct answers in Block 2 compared to Block 1.

To compare the results of Study 3 with of Study 1, we again contrasted the On-Screen Highlighter condition, with the two Gesture conditions combined (*Mean Score*$_{Gesture}$ = 9.26; *SD* = 2.52). As in Study 1, children performed better in the On-Screen Highlighter condition than in the Gesture conditions (see Table 7). Order and Block effects were again significant, indicating that children performed better in the latter condition presented, whichever condition it was, and they performed better in Block 2 compared to Block 1.

Table 5
Study 3: Fixed and random effects for the GLMM predicting children's responses in the robot-led lesson (Reference group: Iconic) (N=37)

Fixed effects	**Estimate**	**SE**	**Wald Z**	***p***
Intercept	0.64	0.43	1.48	.14
Highlighter (vs. Iconic)	0.55	0.26	2.11*	.04
Deictic (vs. Iconic)	0.27	0.34	0.80	.42
Block 2 (vs. Block 1)	0.45	0.22	2.04*	.04
Block 3 (vs. Block 1)	0.37	0.22	1.69	.09
Order 2 (vs. 1)	0.39	0.18	2.13*	.03
Random effects	**Variance**	**SD**		
Subject	1.61	1.27		
Word	0.02	0.14		

* $p < .05$

Table 6
Study 3: Fixed and random effects for the GLMM predicting children's responses in the robot-led lesson (Reference group: Highlighter) (N=37)

Fixed effects	**Estimate**	**SE**	**Wald Z**	***p***
Intercept	1.19	0.4	3.01*	.003
Deictic (vs. Highlighter)	-0.28	0.24	-1.17	.24
Iconic (vs. Highlighter)	-0.55	0.26	-2.11*	.03
Block 2 (vs. Block 1)	0.45	0.22	2.04*	.04
Block 3 (vs. Block 1)	0.37	0.22	1.69	.09
Order 2 (vs. 1)	0.39	0.18	2.13*	.03
Random effects	**Variance**	**SD**		
Subject	1.61	1.27		
Word	0.02	0.14		

* $p < .05$

Table 7.
Study 3: Fixed and random effects for the GLMM predicting children's responses in the robot-led lesson (Highlighter vs. Gesture conditions combined) (N=37)

Fixed effects	**Estimate**	**SE**	**Wald Z**	***p***
Intercept	1.2	0.4	3.02*	.003
Gesture (vs. Highlighter)	-0.41	0.19	-2.17*	.03
Order 2 (vs. 1)	0.39	0.18	2.11*	.04
Block 2 (vs. Block 1)	0.45	0.22	2.04*	.04
Block 3 (vs. Block 1)	0.37	0.22	1.69	.09
Random effects	**Variance**	**SD**		
Subject	1.62	1.27		
Word	0.02	0.15		

* $p < .05$

Further Analyses: Comparing the Human Tutor and Robot Tutor

Both Study 1 and Study 3 revealed no significant difference between iconic and deictic gestures, whether they were performed by a human or a robot tutor. Furthermore, in both studies, children gave more correct answers in the On-Screen Highlighter condition than in the Gesture conditions. However, children performed better in the On-Screen Highlighter condition than in the Deictic condition in Study 1, and the On-Screen Highlighter than in the Iconic condition in Study 3. In order to understand if the tutor type (human vs. robot) makes a difference, we analysed the data from Study 1 and Study 3 together.

As shown in Table 8 and Table 9, Order was significant suggesting that children gave more correct answers in Blocks 2 and 3 than in Block 1. In these models, the three Conditions (Iconic, Deictic, and Highlighter) were not significantly different from one another. However, the interaction between Tutor Type (human vs. robot) and Condition was approaching significance. Figure 2 suggests that children performed better with the robot tutor (Study 3) than with the human tutor (Study 1) in the On-Screen Highlighter and Deictic conditions. However, in the Iconic condition, the two tutors yielded no difference. In other words, children performed equally well with the human and robot tutors when iconic gestures were used in the lessons. The probability of the child giving a correct response in the Iconic condition is comparable in the human and robot studies. Therefore, when the Iconic condition was used as the reference group, the main effect of Tutor Type was largely superseded by two interaction terms - one comparing the Highlighter-Iconic difference of the human study with that of the robot study, and the other comparing the Deictic-Iconic difference of the human study with that of the robot study (Table 8). On the other hand, when the On-Screen Highlighter condition was the reference group, the main effect of Tutor Type remained significant even with the interaction terms (Table 9).

Just as we did in previous analyses, we combined the two Gesture conditions to compare them with the On-Screen Highlighter condition (see Table 10). In addition to the significant effects of Order and Block, the effect of Tutor Type was significant, indicating that children in the robot study outperformed children in the human study.

Table 8
Study 1 vs. Study 3: Fixed and random effects for the GLMM predicting children's responses in human-led and robot-led lessons (Reference group: Iconic) (N=78)

Fixed effects	**Estimate**	**SE**	**Wald Z**	***p***
Intercept	0.57	0.33	1.75	.08
Robot (vs. Human)	0.04	0.39	0.11	.91
Highlighter (vs. Iconic)	-0.13	0.27	-0.47	.64
Deictic (vs. Iconic)	-0.4	0.29	-1.39	.16
Order 2 (vs. 1)	0.39	0.18	2.16*	.03
Block 2 (vs. Block 1)	0.33	0.14	2.30*	.02
Block 3 (vs. Block 1)	0.45	0.14	3.14*	.002
Robot*Iconic-Highlighter (vs. Human*Iconic-Highlighter)	0.68	0.37	1.85	.07
Robot*Iconic-Deictic (vs. Human*Iconic-Deictic)	0.68	0.44	1.54	.12
Random effects	**Variance**	**SD**		
Subject	1.34	1.16		
Word	0.04	0.2		

* $p < .05$

Table 9
Study 1 vs. Study 3: Fixed and random effects for the GLMM predicting children's responses in human-led and robot-led lessons (Reference group: Highlighter) (N=78)

Fixed effects	**Estimate**	**SE**	**Wald Z**	***p***
Intercept	0.45	0.44	1.02	.31
Robot (vsop. Human)	0.72	0.34	2.16*	.03
Deictic (vs. Highlighter)	-0.28	0.29	-0.95	.34
Iconic (vs. Highlighter)	0.13	0.27	0.47	.64
Order 2 (vs. 1)	0.39	0.18	2.16*	.03
Block 2 (vs. Block 1)	0.33	0.14	2.3*	.02
Block 3 (vs. Block 1)	0.45	0.14	3.14*	.002
Robot*Deictic-Highlighter (vs. Human*Deictic-Highlighter)	-0.004	0.37	-0.01	.99
Robot*Iconic-Highlighter (vs. Human*Iconic-Highlighter)	-0.68	0.37	-1.85	.07
Random effects	**Variance**	**SD**		
Subject	1.34	1.16		
Word	0.04	0.20		

* $p < .05$

Table 10
Study 1 vs. Study 3: Fixed and random effects for the GLMM predicting children's responses in the human-led and robot-led lessons (Highlighter vs. Gesture conditions combined) (N=78)

Fixed effects	**Estimate**	**SE**	**Wald Z**	***p***
Intercept	0.45	0.44	1.03	.3
Robot (vs. Human)	0.73	0.34	2.17*	.03
Gesture (vs. Highlighter)	-0.05	0.24	-0.22	.82
Order 2 (vs. 1)	0.39	0.18	2.14*	.03
Block 2 (vs. Block 1)	0.33	0.14	2.30*	.02
Block 3 (vs. Block 1)	0.45	0.14	3.14*	.002
Robot*Gesture-Highlighter (vs. Human*Gesture-Highlighter)	-0.36	0.30	-1.21	.23
Random effects	**Variance**	**SD**		
Subject	1.33	1.15		
Word	0.04	0.20		

* $p < .05$

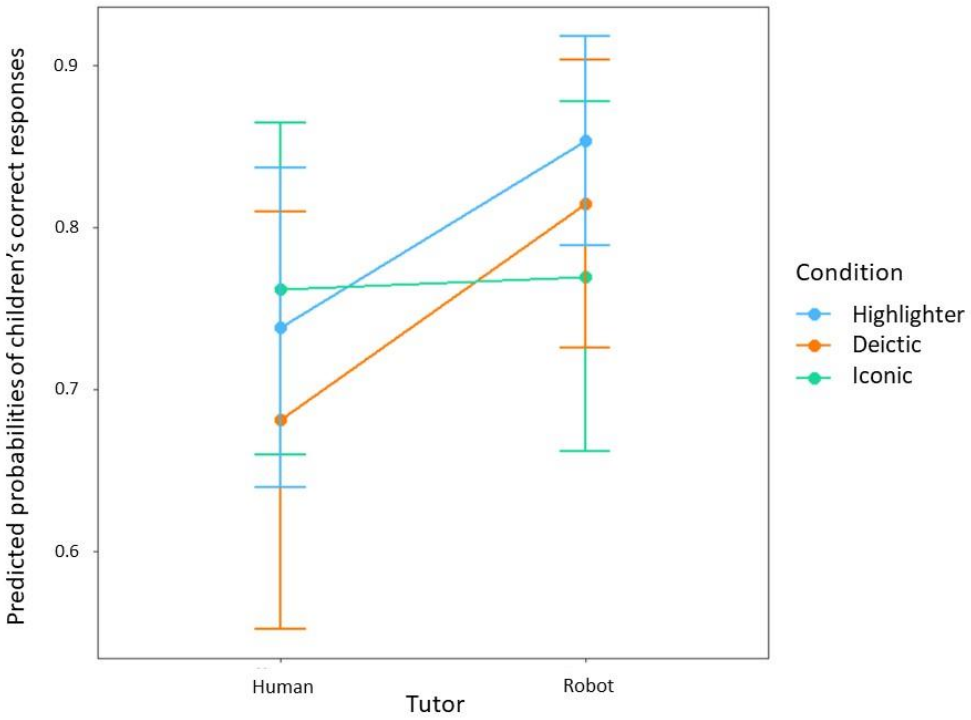

Figure 2. Predicted probabilities of children's correct responses across conditions

Conclusion

Across the three experiments, we investigated whether and how types of gestures affect preschoolers' L2 word learning with human and robot tutors. In Study 1, we found that iconic and deictic gestures performed by a human tutor did not result in different learning outcomes, but that children performed better in the On-Screen Highlighter condition than in the Gesture conditions. In Study 2, we also found no difference between iconic and beat gestures performed by a human tutor. Testing the robot tutor, Study 3 again observed the superiority of the On-Screen Highlighter over Gesture conditions. Additional analyses comparing Study 1 and Study 3 demonstrated that children generally performed better with the robot tutor than with the human tutor. However, there was a trending interaction between Tutor Type and Condition, suggesting that the difference between the On-Screen Highlighter and Iconic conditions may be more pronounced in the robot-led lessons than in the human-led lessons. In other words, especially when the tutor is a robot, drawing attention to the screen where learning material is presented may be more helpful than providing gestures.

3 Robot tutoring of Dutch in Turkish migrant communities

One of the goals of the L2TOR project stated in the revised objectives, was to employ the developed system with Turkish children learning Dutch as a second language. The reason for including this goal, is that one of the potential advantages of using a social robot with immigrant children is that social robots can talk both in the child's first language (L1) and in the language the child is learning (L2). The robot can, therefore, use L1 to support L2 learning. Human teachers of immigrant children do not often speak the children's L1. Moreover, given the diversity seen in immigrant populations currently, it is not likely that a human teacher will be able to support all students in their first language. Some empirical evidence suggests that children learning a second language transfer their skills in L1 into L2 and therefore L1 is supportive for learning of L2 (e.g., Cummins, 1981; Leseman, Henrichs, Blom, & Verhagen, 2017). Cummins (2000) introduced the interdependency hypothesis, suggesting that the more developed children are in their L1, the easier it will be for them to develop L2, given the right conditions (i.e., sufficient exposure for both languages). Thus, transfer from L1 to L2 is not automatic and requires schools and parents to guarantee sufficient exposure to both L2 and L1. Several empirical studies support this hypothesis (e.g., Abu-Rabia, 2001; Hauptman, Mansur, & Tal, 2008). Moreover, the use of L1 in the classroom may also contribute to engagement, self-esteem and positive identity development (Holmes, 2008; Pulinx, van Avermaet, & Agirdag, 2017). Several meta-analyses suggest that indeed, using L1 while teaching L2 is an effective method, though the effects are relatively small and evidence is still scarce (Krashen & McField, 2005; Reljić, Ferring, & Martin, 2014).

The initial results of this study were already reported in deliverable 7.1. Here we report about further analyses done with these data, as we promised in D7.1. Below we provide a short summary of the design of the study. For more details about the design and measures used, see D7.1.

In this study 67 Turkish-Dutch children aged between 48 and 71 months (M = 57.16 months, SD = 6.28) participated. Using a NAO robot and a Microsoft Surface Pro tablet, children were taught six Dutch words for which they knew the Turkish word but no the Dutch word. The study used a within subjects design with two conditions. Half of the target words were taught by a bilingual robot that provided Turkish translations of the target words (L2-L1 condition). The other half of the target words was taught by a monolingual robot that used only Dutch (L2 only condition). All children were presented with both robots, in counterbalanced order. After the lesson, children's knowledge of the target words was measured as well as their vocabulary in both L1 and L2, their enjoyment of the lesson and which of the two robots they preferred. One week after the lesson, knowledge of the target words was again measured.

Previous empirical studies (e.g., Mayo & Leseman, 2008; Demir-Vegter, Aarts, & Kurvers, 2014) have shown that the group of Turkish-Dutch children is extremely heterogeneous in terms of Dutch and Turkish proficiency levels. Therefore, it was not possible to use the same set of target words for all children (as these had to be words the children know in Turkish but not yet in Dutch). Thus, during a pre-test conducted up to one week before the experiment up to 6 target words were chosen for each child separately, out of a list of 20 possible words. See D7.1 for a detailed description of the process of choosing target words and putting together the list of possible target words. Children who had at least four possible target words out of the words included in the study (i.e., words for which they knew the Turkish word but not the Dutch words) were included in the study. About 25% of the children who

participated in the pre-test fewer than four possible target words could be identified. These children were excluded, leaving a final sample size of 67 children.

This design entails that children received different lessons as each child had a different set of words. The post-test included a translation task of the target words and a picture selection task. This picture selection task was also individually constructed per child so that for each trial one picture represented the target word, one distractor a different target word the child was taught and the third picture was a target word the child was not taught.

In D7.1 we reported about initial analyses done with these data. Here we report about further analyses we conducted where we took the words as random factors. This was necessary, as each child was taught a different set of target words. Using this analysis method (linear mixed effects modelling) we can account for possible differences between the target words and, therefore, obtain a clearer picture regarding the differences seen. Moreover, we added the amount of exposure to each target word (dependent on children's performance during the lessons they might have heard target words more often if they required more feedback) as a covariate in the model, in addition to the already included covariates – level of Turkish and Dutch vocabulary and preference of the robot.

All analyses were carried out in R (R Core Team, 2015) using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) where needed. Three separate generalized linear regression models with mixed effects were carried out. To answer our main research question, a basic model was run to investigate the effect of condition (monolingual robot vs. bilingual robot) and time (post-test 1 vs. post-test 2) on children's scores in the post-test, controlling for the number of times a target word had been presented to the child. Our fixed effects were included in this first model, because we were *a priori* interested in their contribution to the outcome (Gelman & Hill, 2007). In a subsequent analysis, Turkish and Dutch vocabulary scores were added as covariates to this model, to explore possible moderation effects. For an exploratory analysis, a final model included children's preference for either the monolingual or bilingual robot as well, to investigate whether children's preference affected learning gains between conditions. In all models, orthogonal sum-to-zero contrast coding was applied to our binary fixed effects (i.e., condition, time, preference) and all continuous variables were centered around zero (Baguley, 2012, p590-621). Furthermore, to avoid problems with non-converging models, we rescaled our continuous variables by dividing them by 10 (Babyak, 2009). We aimed to keep the models as fully specified as possible by including random intercepts for participants and items as well as all within-participant and within-item factors and their possible interactions as random slopes for participant and item (Barr, Levy, Scheepers & Tily, 2013), but because this was not always supported by our relative small data set, we report on the maximal random effect justified by the data (Jaeger, 2010). Finally, to solve issues of non-converging models, we increased the number of possible iterations to 100.000 (Powell, 2009). We report simple rather than standardized effect sizes (Baguley, 2009) and Wald confidence intervals (Agresti & Coull, 1998).

Table 2 displays descriptive statistics for the target word retention scores (post-test scores) for both conditions. Means reflect proportion of correct scores (rather than summed scores) as some participants had missing data and not all participants were tested on the same number of target words. Table 2 also shows the average number of exposures to a target word in Dutch during the lesson. Recall that participants in the bilingual robot condition were exposed to a Turkish translation twice, next to their exposure to the target word in Dutch. This entails, that, while in the L2-L1 condition the number of exposures to

the target words in Dutch is slightly lower, the two conditions had more or less equal number of overall exposures (regardless of language).

Table 2.
Mean proportions correct on the target word retention task for both conditions in the immediate (post-test 1) and delayed post-tests (post-test 2), and average number of exposures to a target word in Dutch.

	L2-only condition			**L2-L1 condition**		
	M	**SD**	**Range**	**M**	**SD**	**Range**
Post-test 1	0.65	0.37	0-1	0.63	0.37	0-1
Post-test 2	0.68	0.38	0-1	0.69	0.36	0-1
Target word exposure in Dutch	10.81	1.69	8-18	8.94	1.65	6-19

Learning gains in L2-only and L2-L1 condition

To investigate the effect of condition and time on children's scores in the post-test, a generalized linear regression model was run, with children's scores on the target word retention task as a dependent variable (0 = incorrect, 1 = correct), condition (L2-only or L2-L1) and time (post-test 1 or post-test 2) as within-participants fixed effects, and the number of exposures as a fixed controlling factor. Condition, time and number of exposures, as well as all their possible interactions, were included as random slopes for participant, because they were within-participant fixed effects. Condition, time and number of exposures, but not their possible interactions, were included as random slope for item, because they were within-participant fixed effects as well.

As shown in Table 3[1], no effect of time was found. Although participants performed better at post-test 2, this effect was not significant (*OR* = 1.42, 95% CI = [.94, 2.15], $z = 1.68$, $p = .093$). There was a main effect of condition, such that participants performed significantly better in the L2-only condition than in the L2-L1 condition (*OR* = 2.25, 95% CI = [1.07, 4.74], $z = 2.14$, $p = .033$). There was also a main effect of target word exposure: performance on the vocabulary task decreased, when the amount of exposures increased (*OR* = .75, 95% CI = [0.60, 0.95], $z = -2.44$, $p = .015$). None of the interactions were significant.

[1] As all continuous variables were rescaled, β-values are not in an interpretable scale either. To get sensible values, one has to divide values from effects with one rescaled variable by 10, values from effects with two rescaled variables by 100 and values from effects with three rescaled variables by 1000. This holds for all three reported models and their outcomes.

Table 3.
Results from the generalized linear regression model with scores from the target word retention as a dependent variable, condition and time as within-participants fixed effects and the number of exposures as a fixed controlling factor.

	B	**SE**	***z***	***p***
Condition	.81	.38	2.14	.033
Time	.35	.21	1.68	.093
Exposures	-2.88	1.18	-2.44	.015
Time * Condition	-.12	.36	-.35	.727
Time * Exposures	-.53	1.16	-.46	.649
Condition * Exposures	.68	1.78	.38	.701
Time * Condition * Exposures	-1.63	2.31	-.71	.480

Examining possible moderation effects of Dutch and Turkish vocabulary skills

Next, we conducted a generalized linear regression model, in which Dutch and Turkish vocabulary scores, as assessed with the Diagnostic Test of Bilingualism, were added as covariates to explore possible moderation effects on the above-reported main effect of condition. In this model, scores on the target word retention (0 or 1) were entered as the dependent variable, condition (L2-only or L2-L1) and time (post-test 1 or post-test 2) as within-participants fixed effects, and the number of exposures, Dutch vocabulary scores, and a Turkish vocabulary score as fixed controlling factors. Condition, time and number of exposures, as well as all their possible interactions, were included as random slopes for participant, because they were within-participant fixed effects. Condition, time and number of exposures, but not their possible interactions, were included as random slopes for item, because they were within-participant fixed effects as well.

As the previous model, this model showed no main effect of time. As above, there was a main effect of condition, indicating that participants performed significantly better with the monolingual robot than with the bilingual robot (*OR* = 2.54, 95% CI = [1.27, 5.08], *z* = 2.63, *p* =.009). Results also showed a main effect of exposures: performance on the vocabulary task decreased, when the amount of exposures increased (*OR* = 0.73, 95% CI = [.58,.92], *z* = -2.69, *p* =.007). Furthermore, a main effect of Turkish vocabulary was found which indicated that performance increased, when Turkish vocabulary scores increased (*OR* = 1.08, 95% CI = [1.03, 1.14], *z* = 3.13, *p* =.002). The results also showed a main effect of Dutch vocabulary, which indicated that performance increased with increasing Dutch vocabulary scores (*OR* = 1.07, 95% CI = [1.01, 1.13], *z* = 2.24, *p* =.025). The interaction effects between condition and either Dutch (*OR* = 1.02, 95% CI = [.91, 1.14], *z* =.35, *p* =.730) or Turkish (*OR* = .98, 95% CI = [.89, 1.09], *z* = -.35, *p* = .728) vocabulary were not significant. Thus, we have no evidence that these vocabulary skills moderate the relation between condition and learning gains. Of all other interactions in the model, only the interaction between Turkish vocabulary, exposures and condition of the robot reached significance (*OR* = .94, 95% CI = [.88, 1.00], *z* = -2.00 *p*= .046). However, because of the large number of comparisons and the rather high p-value of this interaction,

this finding should be treated with caution. For the full results of the model, see Appendix A.

Table 4.
Results from the generalized linear regression model with scores from the target word retention as a dependent variable, condition and time as within-participants fixed effects, and the number of exposures, a Dutch vocabulary score and a Turkish vocabulary score as fixed controlling factors.

	B	**SE**	***Z***	***p***
Condition	0.93	0.35	2.63	0.008
Time	0.17	0.23	0.74	0.458
Dutch vocabulary	0.65	0.29	2.24	0.025
Turkish vocabulary	0.79	0.25	3.13	0.002
Exposures	-3.11	1.16	-2.69	0.007
Condition * Dutch vocabulary	0.19	0.56	0.35	0.730
Condition * Turkish vocabulary	-0.18	0.51	-0.35	0.728

Note. The full model is included in Appendix A.

Children's enjoyment and robot preferences

Self-reported enjoyment was assessed after each condition to investigate possible differences in children's enjoyment between the conditions. However, since these scores showed a ceiling effect (M = 3.84, SD = 0.06 for the monolingual robot, and M = 3.80, SD = 0.06 for the bilingual robot) they were not used in further analyses. None of the participants opted for 'absolutely not enjoyable' and only five participants (both conditions together) opted for 'slightly not enjoyable'. Asking children with which robot they would like to play with again appeared to be more useful to measure children's preference for the monolingual or bilingual robot. Results showed that most of the participants, 48 children (71.6%), preferred to play again with the bilingual robot, and only 19 children (28.4%) stated their preference for the monolingual robot.

To investigate whether children's robot preference affected learning gains differentially between the conditions, preference was added as a between-participant factor in a final generalized linear regression model. As before, this model took scores from the target word retention (0 or 1) as a dependent variable, condition (L2-only or L2-L1) and time (post-test 1 or post-test 2) as within-participants fixed effects, robot preference (preference for monolingual or preference for bilingual) as a between-participants fixed effect, and the number of exposures, a Dutch vocabulary score and a Turkish vocabulary score as fixed controlling factors. Condition, time and number of exposures, but not their possible interactions, were included as random slopes for participant, because they were within-participant fixed effects. Only the number of exposures was included as a random slope for item, as it was a within-participant fixed effect.

Because of the very large number of comparisons in this model, which increases the chance of finding a significant effect, we only report main effects. For the full results of this model, including all interactions, see appendix B. Regarding the main effects, the model showed no effect of time (*OR* = 1.22, 95% CI = [.82, 1.82], *z* = .97, *p* = .330). Again, we found a main effect of condition, as participants performed significantly better with the monolingual robot than with the bilingual robot (*OR* = 2.70, 95% CI = [1.51, 4.85], *z* = 3.33, *p* < .001). Results also showed a main effect of exposures, which indicated that performance on the vocabulary task decreased, when the amount of exposures increased (*OR* = .63, 95% CI = [.50, 0.79], *z* = -3.92 *p* < .001). Furthermore, we found that performance increased, when the Turkish vocabulary knowledge increased (*OR* = 1.09, 95% CI = [1.03, 1.15], *z* = 3.16, *p* = 0.002) and an effect in the same direction was found for Dutch vocabulary (*OR* = 1.10, 95% CI = [1.03, 1.16], *z* = 3.04, *p* = 0.002).

Finally, regarding children's preference for either one of the robots, the results showed that, even though children who had a preference for the monolingual robot performed better than participants who preferred the bilingual robot, this effect was not significant (*OR* = 1.53, 95% CI = [.83, 2.80], *z* = 1.37, *p* = .171).

In sum, across all models we found that children performed better with the monolingual robot than with the bilingual robot. Additionally, children who had more exposures performed worse. Children with larger Turkish and/or Dutch vocabularies performed better during the post-test. Children's preference of the robot was not related to their scores and no interaction effects were found, suggesting that the covariates did not moderate the effect of condition.

The finding that children performed better with the monolingual robot is contrary to our hypothesis. Providing Turkish translations of the target words did not benefit Dutch vocabulary learning.

There are several possible explanations for this finding. Observations of children's reactions during the lessons suggest that children were surprised by the fact that robot purposefully translated the target words. Although children were told that the robot could also speak Turkish, children are possibly not used to the teaching method where the Turkish words are used to support learning the Dutch words. Especially within the school context (our study took place within schools) this method is highly uncommon. Moreover, switching between languages might place extra cognitive load on children. While studies show that bilingual children are usually better at such tasks requiring cognitive flexibility than monolingual children (for a review see: Adesope, Lavin, Thompson, & Ungerleider, 2010), in our study the use of translations does seem to hamper learning. This might be because the use of L1 in this study was minimal and the situation was rather artificial. It might be that a more natural use of L1 within the context of teaching L2, might prove to be more beneficial (see for a discussion: Ticheloven, 2016; for a meta-analysis showing modest effects of bilingual education see: Reljić et al., 2016).

Additionally, we did not measure deep word knowledge in Turkish and Dutch. That is, we only measured receptive knowledge in order to select target words. It is possible that using L1 to support L2 learning is beneficial especially for words where the concept in L1 is already deeply mapped. Future work should attend to this issue. It might also be that the use of L1 is more beneficiary for children with lower levels of L2. While we did see variance in L2 levels in our data, these children were mostly born in the Netherlands and

have some knowledge of Dutch. The use of L1 might be more important for first generation immigrant children with very little knowledge of L2.

Our results also showed that all children enjoyed working with the robots, and a distinct majority had a clear preference for the bilingual Turkish-Dutch robot. This is in line with our hypothesis that the acknowledgement of children's cultural identity through using their L1 in education can increase their enjoyment and wellbeing (Holmes, 2008; Pulinx, van Avermaet & Agirdag, 2017). However, this preference did not affect children's learning gains. This may be due to the fact that there was only little variation in children's robot preferences, which makes it difficult to discover such an effect. Nevertheless, it may possibly increase motivation for future interactions with the robot. More research is needed to address these hypotheses.

4 Teacher feedback during L2 learning

The aim of this study is to investigate how different types of feedback affect children's engagement and learning. In this study, 70 Dutch speaking children aged between five and six years participated. The children were taught 18 different English animal names using the Softbank Robotics NAO robot and a tablet. They played an "I spy with my little eye" game with the robot. The study had a within-subjects design and children participated in three different sessions, which were aimed at teaching six words each. The robot used a different feedback strategy in each session, and the order of feedback strategies and word sets were counterbalanced using a 3x3 latin-square. The three feedback strategies provided by the robot were:

(1) teacher-preferred feedback. For example: 'Well done! You clicked on the horse', 'Too bad, you pressed the bird. Try again! Please click on the horse';
(2) teacher-dispreferred feedback. For example: 'Well done!', 'Too bad';
(3) no feedback.

Only in the teacher-preferred condition, children could try again after they answered the incorrect answer. The robot would repeat the question, but provided help in Dutch to ensure the L2 exposure was the same across conditions. The teacher strategies were based on a survey asking student teachers how they would provide feedback in comparable situations as this study. The lesson was based on the circus experiment described in D5.3.

Our hypotheses are as follows:

(H1) Children will be more engaged (H1a), and will remember more words (H1b) when receiving feedback than receiving no feedback.
(H2) Children will be more engaged (H2a) and will remember more words (H2b) from a robot that provides feedback as preferred by a human teacher than from a robot that provides dis-preferred feedback.

Data collection of this experiment has just been finished and we did not look in detail at the results yet. We are planning to do this in the beginning of January.

5 Gesture production, rating and learning study

Gestures can facilitate language learning in young children (Hostetter, 2011; Sueyoshi & Hardison, 2005; Valenzeno Alibali, & Klatzky, 2003). However, when a given gesture is not appropriate for the context, the gesture does not facilitate, and may even impede, learning. For example, Macedonia, Müller, and Friederici (2011) demonstrated that participants remembered nonce words for a longer period of time when the words were accompanied with iconic gestures than with meaningless gestures (see also Cohen & Otterbein, 1992; Macedonia & von Kriegstein, 2012; So, Chen-Hui, & Wei-Shan, 2012; Tellier, 2008). The facilitatory role of gestures has been recognized in human-robot interaction research, and gestures have been incorporated in language lessons led by a social robot. However, in these studies, the quality of gestures is rarely considered.

Gesture production, rating, and learning study aimed to identify gestures that are suitable for teaching our target words and to evaluate whether the match between words and gestures (i.e., how well the gesture represents the corresponding word) predicts learning outcomes. Asking this question is especially important for our project because gestures can be counterproductive when performed by the NAO, (1) whose movements are not as smooth as human gestures, and (2) who produces fairly loud motor sounds in performing gestures and other actions. Importantly, previous research on human gestures (e.g., Macedonia et al., 2011) compared iconic gestures with non-iconic or meaningless gestures, and thus does not inform us if iconic gestures can facilitate word learning as long as they are "good enough" or if they need to be in very good quality. Therefore, this project used iconic gestures all of which represented our target words fairly well, but still differed in how well they represented the words.

The project consisted of three studies: production study, rating study, and learning study. The *production* and *rating studies* were conducted to prepare stimuli used in the *learning study*. We first asked adult English native speakers (N = 3) to produce gestures corresponding to our target words and filmed their gestures (the production study). Another group of adult English native speakers (N = 20) were asked to rate how well those gestures represented the words on a scale of 1-7 (the rating study). The NAO was manually animated to perform gestures that are as similar as possible to the human gestures (Figure 1).

Based on the production and rating studies, we selected five word-gesture pairs that varied in their rating scores (*sliding*: 3.56, *falling*: 4.72, *climbing*: 5.92, *walking*: 6.14, and *throwing*: 6.28) to be used in an experiment with children, *the learning study*. Originally, *jumping* was also included on the list but later excluded because it was impossible for the NAO to perform the movement (actual jumping in which the gesturer jumps off the ground) and even the closest movement (bending the knees and extending them quickly) overloaded the NAO's knee motors when performed multiple times in a short period of time. We selected words of the same part of speech, i.e., verb, to minimize the difference in the meanings of the target words.

Figure 1. One of the gestures produced in the production study (left), and the corresponding gesture performed by the NAO (right). The gesture was for the word "sliding."

The learning study was conducted in two conditions: Robot Tutor and Human Tutor. In both conditions, children learned the five target words in one-on-one lesson with a tutor. We did not use any other devices such as a laptop in the lesson because our Gesture Type Study (see *2. Gesture type study: how do human tutors gesture? – an update of the results* in this deliverable) suggested that having another device may lead children to focus too much on the device and not the robot. We, however, used a laptop in the tests of receptive vocabulary, which required the presentation of animations. In addition to the robot or human tutor, another experimenter, or the *human tester*, was present in the room and administered the productive and receptive vocabulary tests (Figure 2; see below for the details of the vocabulary tests).

Figure 2. Experimental setting in the learning study. The child received the L2 lesson from the tutor (the NAO or a human adult) and was tested in the productive and receptive tests administered by another human experimenter, i.e., the human tester (bottom left of the picture).

The data reported here include 21 children in the Robot condition (*Mean age* = 65.81 months; *SD* = 6.87), and 22 children in the Human condition (*Mean age* = 72.32

months; SD = 6.76). Additional 3 children were tested in the Robot condition, but were excluded from the analysis as they did not complete the task. A session consisted of three sections. The first was two pre-lesson tests. In *the receptive test*, the child was presented with two animations on the laptop screen and was asked to point to the animation that corresponded to each of the five verbs. *The productive test* was essentially a translation task in which the child was asked to say the meaning of each target word in Turkish. The second section of the session was three blocks of a lesson and a receptive test. In the lesson part, the tutor (a robot or a human adult) taught the five target words to the child using the iconic gestures described above. After each lesson, the receptive test was administered by the human tester described above.

To examine whether Word-Gesture Match (i.e., word-gesture match ratings from the rating study) predicts accuracy in the receptive tests, we analysed the data using Generalized Linear Mixed Models (GLMMs) with logit (log-odds) as the link function as we did for Gesture Type Study (*2. Gesture type study: how do human tutors gesture? – an update of the results* in this deliverable). GLMMs with logit as the link function are essentially logistic regressions with both fixed and random effects, which allow us to analyse the accuracy data without averaging across trials. All GLMMs were generated in R (R Development Core Team, 2016) using the *lme4.glmer* function (Bates, Maechler, Bolker, & Walker, 2014). We used the *lmerTest* package to obtain p values for the fixed effects (Kuznetsova, Brockhoff, & Christensen, 2017).

Table 1 summarizes the results of the productive tests. Except for the words "falling" and "throwing" in the Human Tutor condition, the number of children who correctly produced the Turkish translations of the target English words increased from the pre-lesson to the post-lesson. The results suggest that the lesson successfully increased children's vocabulary. However, we found no systematic pattern regarding which words are learned better than others. We built a GLMM including a fixed intercept, fixed effects for Word-Gesture Match (i.e., word-gesture match ratings from the rating study), Tutor Types (Robot vs. Human), Pre-Lesson Productive Test Scores, and a random intercept for subjects (Table 2). The model found Tutor Type to be significant, suggesting that children in the Robot Tutor condition performed better in the productive test than did children in the Human Tutor condition. Age of children was also marginally significant, indicating that older children may have performed better than younger children.

Table 1

Percentages of children who correctly produced the Turkish translations of the target English words in the productive tests before ("pre-lesson") and after ("post-lesson") the language lesson led by the robot or human tutor. Words are aligned from the lowest (sliding) to the highest (throwing) word-gesture match ratings.

		Sliding	Falling	Climbing	Walking	Throwing
Robot	Pre-Lesson	10%	20%	10%	10%	5%
	Post-Lesson	38%	38%	14%	33%	19%
Human	Pre-Lesson	5%	5%	5%	0%	14%
	Post-Lesson	23%	5%	14%	18%	14%

Table 2
Fixed effects for the GLMM model predicting the accuracy in the post-lesson productive test.

	B	*SE*	Wald Z	*p*
(Intercept)	-4.22	2.30	-1.84 +	.07
W-G Match	-0.26	0.16	-1.60	.11
Tutor Type	1.20	0.43	2.81 **	.00
Age	0.05	0.03	1.82 +	.07
Pre-Lesson Accuracy	-1.51	1.07	-1.41	.16

$^{+}p < .10$. $^{**}p < .01$.

The receptive tests resulted in a different pattern. Figure 3 shows the accuracy in the three receptive tests administered after each lesson. Visual inspection of the data suggests that, especially in the Human Tutor condition, the match between gesture and word may have affected learning outcomes. Confirming the idea, a GLMM predicting the accuracy in the receptive tests found the fixed effects of Word-Gesture Match to be significant (Table 3). In addition to the factors included in the model for the productive test, this model included Block (Block 1 vs. Block 2 vs. Block 3) and the interaction between Word-Gesture Match (i.e., word-gesture match ratings from the rating study), and Tutor Type (Robot vs. Human). This model also suggests that children in the Robot Tutor condition performed better in the receptive tests, and older children performed better than younger children. However, the interaction between Word-Gesture Match and Tutor Type did not reach statistical significance.

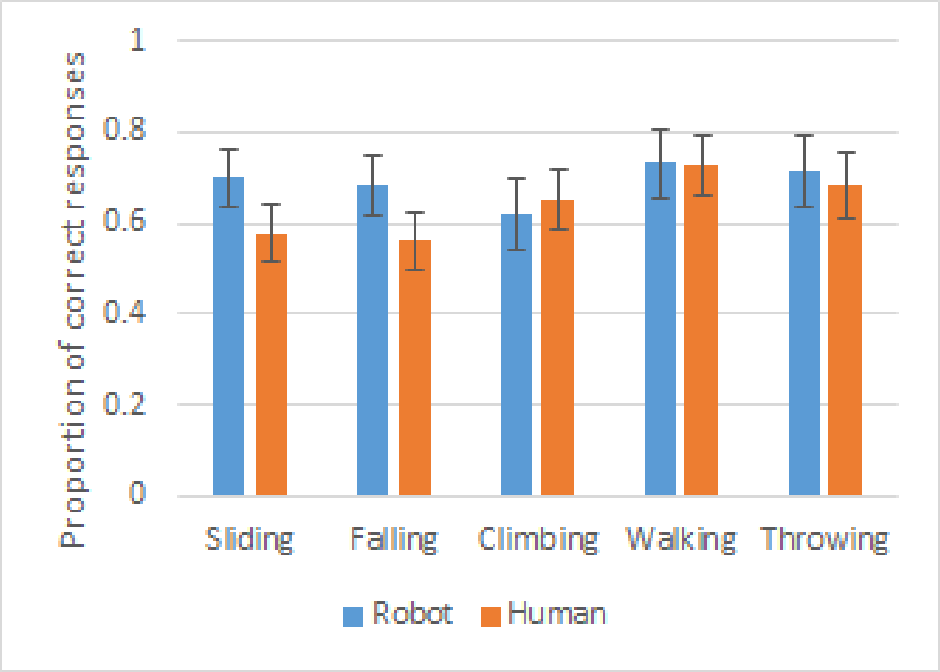

Figure 3. Accuracy in the three receptive tests administered after each lesson. Words are aligned from the lowest (sliding) to the highest (throwing) word-gesture match ratings.

Table 3
Fixed effects for the GLMM model predicting the accuracy in the three receptive tests administered after each lesson.

	B	*SE*	Wald Z	*p*
(Intercept)	-4.57	1.43	-3.20 **	<.01
W-G Match	0.24	0.11	2.09 *	.04
Tutor Type	1.82	0.92	1.98 *	.05
Block	0.15	0.11	1.38	.17
Age	0.05	0.02	2.87 **	<.01
Pre-Lesson Test	0.03	0.19	0.14	.89
W-G Match x Tutor Type	-0.24	0.17	-1.43	.15

*p < .05. **p < .01.

In summary, we suggest that the match between words and gestures can affect how well children learn L2 words. Thus, iconic gestures should be evaluated and validated properly before implemented in robot-assisted L2 lessons. This pattern was only found in the receptive tests and not in the productive test. It is safe to say that the results of the receptive tests represent children's learning outcomes better than that of the productive test, because receptive tests are generally a more sensitive measure of children's vocabulary knowledge than productive tests (e.g., Webb, 2008). We also tested children three times on receptive vocabulary to acquire reliable data. Interestingly, this study also indicates that children in the Robot Tutor condition learned better than children in the Human Tutor condition. While the reason behind the difference is unclear, the pattern aligns with the results of Gesture Type Study (see *2. Gesture type study: how do human tutors gesture? – an update of the results* in this deliverable). Our data also suggest that the match between words and gestures may be more pronounced when a human adult is the tutor than when the NAO is the tutor though the interaction term did not reach statistical significance. Our sample size is not very large, and thus we plan to test more participants in the future to evaluate the possibility of the gesture quality differently affecting learning outcomes in the robot- and human-led L2 lessons.

6 Influence of robot platform on learning

An often asked question is what the influence is of the robot's appearance, or the *robot platform*, on learning. Are taller robots more effective than small robots? Is a more humanoid robot better? The question is not only of relevance to researchers studying robots for learning, but is very relevant for industry as well. Industry has an interest in keeping the complexity, and therefore the cost, of a product as low as possible. This not only has an impact on the selling price of products, but also has a strong influence on after sales services.
We tried to answer this question not by running an experiment comparing two different robot platforms, but instead used the meta-analysis reported in (Belpaeme et al., 2018) to get a view on what the contribution is of the robot platform on learning outcomes. Belpaeme et al. (2018) report on 101 published papers, containing over 300 study results. In the meta-analysis a wide range of robot platforms is used to study robots for learning (see figure 1).

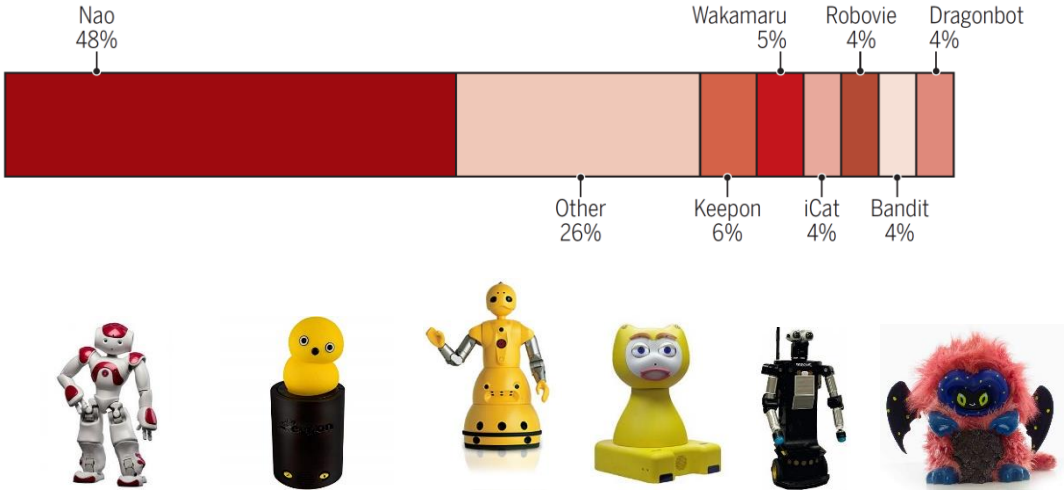

Figure 1: robots featured in the meta-analysis, from left to right: Nao, Keepon, Wakamaru, iCat, Robovie and Dragonbot.

While studies all differ in the populations used, the study design and the subjects tutored by the robot, there is sufficient data on two different robot platforms to make an informative comparison. The most popular robot in the studies we analysed is the Nao robot, a 54-cm-tall humanoid by Softbank Robotics Europe available as having 14, 21, or 25 degrees of freedom. The two latter versions of Nao have arms, legs, a torso, and a head. They can walk, gesture, and pan and tilt their head. Nao has a rich sensor suite and an on-board computational core, allowing the robot to be fully autonomous. The dominance of Nao for HRI can be attributed to its wide availability, appealing appearance, accessible price point, technical robustness, and ease of programming. Hence, Nao has become an almost de facto platform for many studies in robots for learning. Another robot popular as a tutor is the Keepon robot, a consumer-grade version of the Keepon Pro research robot. Keepon is a 25-cm-tall snowman-shaped robot with a yellow foam exterior without arms

and legs. It has four degrees of freedom to make it pan, roll, tilt, and bop. Originally sold as a novelty for children, it can be used as a research platform after some modification. Nao and Keepon offer two extremes in the design space of social robots, and hence, it is interesting to compare learning outcomes for both.

Comparing Keepon with Nao, the respective cognitive learning gain is $d = 0.56$ ($N = 10$; 95% CI, 0.532 to 0.58) and $d = 0.76$ ($N = 8$; 95% CI, 0.52 to 1.01); therefore, both show a medium-sized effect. However, we note that direct comparisons between different robots are difficult with the available data, because no studies used the same experimental design, the same curriculum, and the same student population with multiple robots. Furthermore, different robots have tended to be used at different times, becoming popular in studies when that particular hardware model was first made available and decreasing in usage over time. Because the complexity of the experimental protocols has tended to increase, direct comparison is not possible at this point in time.

What is clear from surveying the different robot types is that all robots have a distinctly social character. All robots have humanoid features—such as a head, eyes, a mouth, arms, or legs—setting the expectation that the robot has the ability to engage on a social level. Although there are no data on whether the social appearance of the robot is a requirement for effective tutoring, there is evidence that the social and agentic nature of the robots promotes secondary responses conducive to learning. The choice of robot very often depends on practical considerations and whether the learners feel comfortable around the robot. The weighted average height of the robots is 62 cm; the shortest robot in use is the Keepon at 25 cm, and the tallest is the RoboThespian humanoid at 175 cm. Shorter robots are often preferred when teaching young children.

7 Children's reliance on the non-verbal cues of a robot

Introduction

An often implicit assumption in robot-assisted language learning (RALL) studies is that learners can employ the non-verbal behaviors of a robot, such as eye gaze and pointing, for learning. Young children rely on non-verbal behaviors to learn new label-meaning mappings (Baldwin, 1991; Baldwin et al, 1996) and disambiguate between possible meanings of new words (Brojde, Ahmed, & Colunga, 2012; Grassmann & Tomasello, 2010; Meyer & Baldwin, 2013) if these are provided by a human. However, it is as yet unknown whether children rely to the same extent on non-verbal cues if these are provided by a robot. The primary aim of the current study is to investigate whether children rely on a robot's pointing and eye gaze if these are contrasted with verbal labels to the same extent as with a human speaker. A further aim of the study is to see whether children's reliance on non-verbal cues of a robot is related to their perception of the robot as a human-like entity, that is, to the degree to which they anthropomorphize the robot.

A number of studies have demonstrated that children rely more strongly on a non-verbal cue than on a verbal cue in figuring out which object a speaker refers to. Grassmann and Tomasello (2010) administered a disambiguation task in which children were presented with two objects (e.g., a car and a novel object). The experimenter then verbally referred to one of these objects ("Give me the car"), while she pointed at the novel object, or vice versa (i.e., the experimenter asked for "toma", while pointing at the car). Grasmann and Tomasello found that German two- and four-year-old children relied on pointing more strongly than on labeling in resolving this conflict, as children overwhelmingly handed the object pointed at to the experimenter. This preference for pointing was stronger when the experimenter used a novel label (e.g., "modi") while pointing at a familiar object than when she used a familiar label (e.g., "car") while pointing at an unfamiliar object. On the basis of these findings, the authors concluded that young children attribute more importance to socio-pragmatic cues than to verbal cues when resolving a referential conflict, especially so if they are uncertain about the meaning of a word (for replication studies, see Ateş, 2016; Grassmann, Magister, & Tomasello, 2011; Verhagen, Grassmann, & Küntay, 2017). The current study addresses three questions:

(1) How do children weigh non-verbal cues (i.e., eye gaze and pointing) and verbal cues (i.e., labeling) from a robot versus a human speaker?
(2) Do children weigh such non-verbal cues differently depending on whether these are contrasted with a novel label or a familiar label? Do any effects of label familiarity differ between a robot and a human?
(3) Do children rely differently on a robot's non-verbal versus verbal cues depending on the degree to which they anthropomorphize the robot?

We report on two studies that were conducted to address these questions. In Study 1, we tested children's reliance on eye gaze versus labeling. In Study 2, we tested children's reliance on pointing versus labeling. In both studies, children's following of the non-verbal cue versus the labeling cue was compared across two conditions: one in which a robot provided the cues and one in which a human adult provided these cues. In each study, non-verbal cues were contrasted with a verbal cue that either involved a familiar

label (e.g., "car") or a novel label (e.g., "modi"), following earlier work (Grassmann & Tomasello, 2010; Verhagen et al., 2017).

Study 1

In Study 1, we investigated children's reliance on eye gaze versus a verbal label in a disambiguation task that was either administered by a social robot or a human speaker. Participants were 42 monolingual Dutch children (25 girls, 60%) with an average age of 60 months (*SD* = 6, range = 50 – 74). In the disambiguation task, modeled after the task reported in Grassmann and Tomasello (2010) and Verhagen et al. (2017), a referential conflict was created by pitting a non-verbal (gazing) cue and a verbal (labeling) cue against each other. Two conditions were tested. First, in the 'familiar label condition', the experimenter said the Dutch equivalent of the following instruction "Let's play with the car. Tap on the car". While producing this instruction, she gazed at the novel object. In the 'novel label' condition, the experimenter said the Dutch equivalent of "Let's play with the modi. Tap on the modi". While producing this, she gazed at the familiar object (i.e., car). The experiment had a 2x2 design. Besides the 'label' condition (i.e., novel label vs. familiar label), there was a 'speaker' condition, as the task was either administered by a robot or by a human. Both the 'label' and 'speaker' conditions were administered within-subjects, so that each child was presented with the robot and the human, and performed both the familiar label and novel label trials. The two 'speaker' conditions were administered in two different sessions that were on average one week apart. In addition, we used a questionnaire adapted from Jipson and Gelman (2007) to assess to what extent children perceived the robot as a human-like entity. It contained twelve yes/no-questions, as well as, for each question, the follow-up question "Why?" or "Why not?". Example questions are "Can Robin the robot see things?", "Should Robin the robot eat?", and "Can Robin the robot be happy?".

Linear mixed-effects models indicated that children followed eye gaze significantly below chance in all conditions (i.e., $t(39) = -5.019$, $p < .001$, $d = 1.29$ for the robot using a novel label; $t(40) = -12.858$, $p < .001$, $d = 2.01$ for the robot using a familiar label; $t(39) = -4.286$, $p < .001$, $d = 0.68$ for the human using a novel label; $t(39) = -9.635$, $p < .001$, $d = 1.52$ for the human using a familiar label). They thus overwhelmingly relied on the verbal label instead, irrespective of whether a robot or a human administered the task. Our results also showed that children's relied on gaze more strongly when the gaze cue was contrasted with a novel label than with a familiar label, $ß = 2.45$, $SE = .45$, $z = 5.40$, $p < .001$. The degree to which children perceived of the robot as resembling a human did not predict children's gaze following, $ß = -.15$, $SE = .23$, $z = -.66$, $p = .51$.

Study 2

The aims of Study 2 were similar to those of Study 1, as we investigated whether (i) children's reliance on pointing versus labeling differed between a robot and a human speaker, (ii) children showed a smaller effect of label familiarity with a robot versus a human, and (iii) children who considered the robot as human-like relied more strongly on its pointing gestures than children who considered it less human-like. Participants were 60 monolingual Dutch kindergartners (22 girls, 37%) with a mean age of 62 months (SD = 6, range = 50 – 74). The design of the task was the same was in Study 1, except that the

experimenter pointed at rather than gazed at one of the images, while verbally labeling the other image (as in Grassmann & Tomasello, 2010; Verhagen et al., 2017).

Linear mixed-effects models indicated, first, that children followed the pointing gesture significantly above chance in all conditions (i.e., $t(54) = 2.869$, $p = .006$, $d = 0.39$ for the robot using a novel label; $t(54) = 3.125$, $p = .003$, $d = 0.43$ for the robot using a familiar label; $t(55) = 2.195$, $p = .032$, $d = 0.29$ for the human speaker using a novel label; $t(55) = 2.445$, $p = .018$, $d = 0.33$ for the human speaker using a familiar label), in keeping with previous studies (Grassmann & Tomasello, 2010; Verhagen et al., 2017; Magister, Grassmann, & Tomasello, 2011). However, unlike in these earlier studies, no differences were observed in children's point following depending on whether pointing was pitted against a familiar or a novel label, $ß = -.23$, $SE = .25$, $z = -.94$, $p = .261$.

Yet, when children's perception of the robot as displaying human-like properties was taken into account, a significant interaction between 'perception' and 'label' emerged, $ß = .40$, $SE = .09$, $z = 4.46$, $p < .001$. Children who perceived of the robot as displaying many human-like properties relied on pointing more after hearing a novel label rather than a familiar label, while children who perceived of the robot as being little human-like relied on pointing more after hearing a familiar label rather than a novel label. A three-way interaction between 'speaker', 'label' and 'perception' showed a trend towards significance, moreover, $ß = -.30$, $SE = .17$, $z = -1.71$, $p = .088$, indicating that the differential effect of label familiarity for children varying in perception scores mainly held for the robot condition.

General Discussion

In this study, we investigated how children weighed non-verbal communicative cues (i.e., eye gaze and pointing) and verbal cues (labelling) of a robot as compared to those of a human. In two studies, children's reliance on non-verbal cues was assessed, using disambiguation tasks in which a robot or a human presented a conflict between a non-verbal and a verbal cue. The verbal cue either involved a familiar verbal label (e.g., "car") or an unfamiliar verbal label (e.g., "modi").

Our results showed that children did not differ in their following of non-verbal and verbal cues in resolving the conflict between a robot and a human. This held true regardless of whether eye gaze or pointing was used. Effects of label familiarity were found in both studies. In Study 1, children relied more strongly on eye gaze when it contrasted with a novel verbal label than when it contrasted with a familiar label (both with a robot and a human). In Study 2, children followed the non-verbal (pointing) cue more often if it contrasted with a novel label than with a familiar label, but this difference was only found for children who considered the robot as human-like, as children who considered the robot as less human-like showed the opposite effect. However, this interaction should be interpreted with caution, given that it was only slightly stronger in the robot than in the human condition.

Our results are in keeping with earlier work showing that children do not differ in their following of non-verbal cues between a robot and a human (Kory Westlund et al., 2017). In our study, children relied much less strongly on eye gaze than on pointing, in line with earlier results for three- and four-year-olds and a human experimenter (Jaswal & Hansen, 2006), and previous results for toddlers' reliance on eye gaze in a similar task (Graham et al., 2010).

Our main finding that children did not differ in their reliance on non-verbal as opposed to verbal cues from a robot versus a human has important implications for RALL studies. Crucially, it opens up possibilities for designing educational programs in which robots use non-verbal communicative cues to support children's learning. However, more research into this topic is needed. While previous RALL studies have looked into the added value of robots' use of (iconic) gestures (de Wit et al., 2018) or gesturing as part of the robot's tutoring program (Alemi, Meghdari, & Ghazisaedy, 2014), to the best of our knowledge, no earlier studies have investigated whether a robot's use of eye gaze or pointing positively affect children's learning. Also, future research could address in more detail how children's anthropomorphism relates to how children interact with robots, and on children's learning outcomes. Differences in anthropomorphism are not trivial, as they may be related to children's trust in a robot and, in turn, to the socio-emotional relationships they may or may not establish with a robot. As such, they bear on important ethical issues not to be neglected in child-robot interaction research.

8 Reference List

Abu-Rabia, S. (2001). Testing the interdependence hypothesis among native adult bilingual Russian-English students. *Journal of Psycholinguistic Research, 30,* 437–455. doi:10.1023/A:1010425825251

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research, 80*(2), 207-245. doi: 10.3102/0034654310368803.

Agrest, A., & Coull, B. A. (1998). Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician, 52*(2), 119-126.

Babyak, M. A. (2009). Post to Statistical Tips from the Editors of *Psychosomatic Medicine*. September 24, 2009. Retrieved from http://stattips.blogspot.com/2009/08/rescaling-continuous-predictors-in.html

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology,* 100, 603-617.

Baguley, T. (2012). *Serious Stats*. Basingstoke, UK: Palgrave Macmillan.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed- Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7 [Computer software]. Retrieved from http://CRAN.R-project.org/package=lme4

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science Robotics*, 3(21), eaat5954.

Cohen, Ronald L. & Nicola Otterbein (1992). The mnemonic Effect of speech Gestures: Pantomimic and Non-Pantomimic Gestures compared. European Journal of Cognitive *Psychology*, *4*(2), 113-139.

Cummins, J. (1981). Bilingualism and minority language children. Ontario: Ontario Institute for Studies in Education.

Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Clevedon, England: Multilingual Matters.

Demir-Vegter, S., Aarts, R., & Kurvers, J. (2014). Lexical richness in maternal input and vocabulary development of Turkish preschoolers in the Netherlands. *Journal of Psycholinguistic Research, 43*(2), 149-165. doi:10.1007/s10936-013-9245-7

Gelman, A., & J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge, MA: Cambridge University Press.

Hauptman, S., Mansur, F., & Tal, R. (2008). A trilingual teaching model for developing academic literacy skill in classical Arabic (L1), Hebrew (L2) and English (FL) in

Southern Israel. *Journal of Multilingual and Multicultural Development, 28*, 181–197. doi:10.2167/jmmd530.0

Holmes, J. (2008). *An introduction to sociolinguistics* (3rd ed.). England: Pearson Longman.

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological bulletin*, *137*(2), 297-315.

Jaeger, T. F. Post to HLP/Jaeger lab blog. May 14 2009. Retrieved from https://hlplab.wordpress.com/2009/05/14/random-effect-structure/

Krashen, S., & McField, G. (2005). What works? Reviewing the latest evidence on bilingual education. *Language Learner, 1*(2), 7–10.

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1-26.

Leseman, P. P. M., Henrichs, L. F., Blom, E., & Verhagen, J. (2017). Young mono- and bilingual children's exposure to academic language as related to language development and school achievement. In V. Grøver, P. Ucelli, M. Rowe & E. Lieven (Eds.), *Learning through language*. Cambridge, MA: Cambridge University Press.

Macedonia, M., & von Kriegstein, K. (2012). Gestures enhance foreign language learning. *Biolinguistics*, *6*(3-4), 393-416.

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.

Mayo, A. Y., & Leseman, P. P. M. (2008). Off to a Good Start? Vocabulary development and characteristics of early family and classroom experiences of children from native Dutch speaking and bilingual minority families in the Netherlands. *Educational andChild Psychology*, *25*(3), 66-78.

Powell, M. J. D. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives*. Technical Report, Department of Applied Mathematics and Theoretical Physics, University of Cambridge.

Pulinx, R., Van Avermaet, P., & Agirdag, O. (2017). Silencing linguistic diversity: The extent, the determinants and consequences of the monolingual beliefs of Flemish teachers. *International Journal of Bilingual Education and Bilingualism*,*20*(5), 542-556. https://doi.org/10.1080/13670050.2015.1102860

R Core Team. 2015. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

R Development Core Team. (2016). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Reljić, G., Ferring, D., & Martin, R. (2014). A Meta-Analysis on the Effectiveness of Bilingual Programs in Europe. *Review of Educational Research, 85*, 92-128. doi: 10.3102/0034654314548514.

So, W. C., Chen-Hui, C. S., & Wei-Shan, J. L. (2012). Mnemonic effect of iconic gesture and beat gesture in adults and children: Is meaning in gesture important for memory recall? *Language and Cognitive Processes*, *27*(5), 665-681.

Sueyoshi, A., & Hardison, D. M. (2005). The Role of Gestures and Facial Cues in Second Language Listening Comprehension. Language Learning, 55(4), 661-699.

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

Ticheloven, A. (2016). *Translanguaging as pedagogy in a superdiversity classroom: Constraints and opportunities.* Unpublished Master's thesis, Utrecht University, Utrecht, The Netherlands.

Valenzeno, L., Martha Alibali, and Roberta Klatzky. 2003. Teachers' Gestures Facilitate Students' Learning: A Lesson in Symmetry. Contemporary Educational Psychology 28:187-204.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second language acquisition*, *30*(1), 79-95.

Appendix A: Models with markdown for Robot tutoring of Dutch in Turkish migrant communities

This markdown contains the analyses for the study: Robot tutoring of Dutch in Turkish migrant communities

General comment:

We aimed to keep the models as fully specified as possible by including random intercepts for participants and items as well as all within-participant and within-item factors and their possible interactions as random slopes for participant and item (Barr, Levy, Scheepers & Tily, 2013), but because this was not always supported by our relative small data set, we run the maximal random effect structures justified by the data (Jaeger, 2010). To solve issues of non-converging models, we increased the number of possible iterations to 100.000 (Powell, 2009). We calculated Wald confidence intervals (Agresti & Coull, 1998).

Load required packages

```
options(width = 150)
library (lme4)
## Warning: package 'lme4' was built under R version 3.2.5
## Loading required package: Matrix
```

Load data

```
table <- read.csv("rawData.csv", sep=",")
table [1:10, 1:9]
##      X Subject     Item Age  Time Condition Score Pref DutchV
## 1    1     101 bibberen  63 Test1    L2Only     1   Bi       9
## 2    2     101 bibberen  63 Test1    L2Only     1   Bi       9
## 3    3     101 bibberen  63 Test1    L2Only     1   Bi       9
## 4    4     101 bibberen  63 Test2    L2Only     1   Bi       9
## 5    5     101 bibberen  63 Test2    L2Only     1   Bi       9
## 6    6     101 bibberen  63 Test2    L2Only     1   Bi       9
## 7    7     101    legen  63 Test2    L2Only     1   Bi       9
## 8    8     101    legen  63 Test2    L2Only     1   Bi       9
## 9    9     101    legen  63 Test1    L2Only     1   Bi       9
## 10 10     101    legen  63 Test2    L2Only     1   Bi       9
```

Ensure subject is regarded as a factor

```
table$Subject <- as.factor(table$Subject)
```

Apply orthogonal sum-to-zero contrast coding to the binary fixed effect ‘time’ (Baguley, 2012, p590-621)

```
contrast <- cbind (c(-1/2, +1/2))
colnames (contrast) <- c("-T1+T2")
contrasts (table$Time) <- contrast
```

Apply orthogonal sum-to-zero contrast coding to the binary fixed effect ‘condition’ (Baguley, 2012, p590-621)

```
contrast <- cbind (c(-1/2, +1/2))
colnames (contrast) <- c("-L2L1+L2Only")
contrasts (table$Condition) <- contrast
```

Apply orthogonal sum-to-zero contrast coding to the binary fixed effect ‘preference’, for the exloratory analysis, reported in model 3 (Baguley, 2012, p590-621)

```
contrast <- cbind (c(-1/2, +1/2))
colnames (contrast) <- c("-Bi+Mono")
contrasts (table$Pref) <- contrast
```

Center en rescale the continuous variable ‘exposure’ (Baguley, 2012, p590-621; Babyak, 2009)

```
Exposure <- aggregate(Exposures ~ Subject, table, mean)
table$Exposures <- table$Exposures - mean(Exposure$Exposures)
table$Exposures <- table$Exposures/10
```

Center en rescale the vocabulary scores (Baguley, 2012, p590-621; Babyak, 2009)

```
DutchV <- aggregate(DutchV ~ Subject, table, mean)
table$DutchV <- table$DutchV - mean(DutchV$DutchV)
table$DutchV <- table$DutchV/10
TurkishV<- aggregate(TurkishV ~ Subject, table, mean)
table$TurkishV <- table$TurkishV - mean(TurkishV$TurkishV)
table$TurkishV <- table$TurkishV/10
```

Learning gains in L2-only and L2-L1 condition

To investigate the effect of time and condition on children's scores in the post-test, a generalized linear regression model was run, with children's scores on the target word retention task as a dependent variable (0 = incorrect, 1 = incorrect), condition (L2-only or L2-L1) and time (post-test 1 or post-test 2) as within-participants fixed effects and the number of exposures as a fixed controlling factor. Condition, time and number of exposures, as well as all their possible interactions, were included as random slopes for participant, because they were within-participant fixed effects. Condition, time and number of exposures, but not their possible interactions, were included as random slope for item, because they were within-participant fixed effects as well.

```
model <- glmer(Score ~ Time * Condition * Exposures + (Exposures * Condit
ion * Time | Subject) + (Exposures + Condition + Time | Item), data=table
, family="binomial", REML=FALSE, glmerControl(optimizer="bobyqa", optCtrl
 = list(maxfun = 100000)))

summary (model)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace Appr
oximation) ['glmerMod']

##  Family: binomial  ( logit )

## Formula: Score ~ Time * Condition * Exposures + (Exposures * Condition
 *      Time | Subject) + (Exposures + Condition + Time | Item)

##    Data: table

## Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e
+05))

##

##      AIC      BIC   logLik deviance df.resid

##   2279.1   2582.6  -1085.6   2171.1     1984

##

## Scaled residuals:

##     Min      1Q  Median      3Q     Max

## -5.4673 -0.6622  0.2694  0.5450  2.9000

##

## Random effects:

##  Groups  Name                                        Variance Std.Dev.
Corr

##  Subject (Intercept)                                  1.18491 1.0885

##          Exposures                                   24.57226 4.9570
-0.28
```

```
##           Condition-L2L1+L2Only                          2.24636 1.4988
 0.42 -0.50

##           Time-T1+T2                                     0.30104 0.5487
 0.23  0.27  0.39

##           Exposures:Condition-L2L1+L2Only                5.66871 2.3809
-0.42 -0.20 -0.12 -0.89

##           Exposures:Time-T1+T2                           9.02753 3.0046
-0.87 -0.04 -0.43 -0.67  0.76

##           Condition-L2L1+L2Only:Time-T1+T2               0.65614 0.8100
 0.48 -0.32  0.83  0.79 -0.65 -0.69

##           Exposures:Condition-L2L1+L2Only:Time-T1+T2 43.01150 6.5583
-0.42 -0.59  0.06 -0.78  0.87  0.77 -0.39

##  Item     (Intercept)                                    1.08367 1.0410


##           Exposures                                      0.54582 0.7388
-0.72

##           Condition-L2L1+L2Only                          0.58017 0.7617
 0.28  0.47

##           Time-T1+T2                                     0.03149 0.1775
-0.21 -0.53 -1.00

## Number of obs: 2038, groups:  Subject, 67; Item, 19

##

## Fixed effects:

##                                             Estimate Std. Error z value
 Pr(>|z|)

## (Intercept)                                   1.3510     0.3089   4.374
 1.22e-05 ***

## Time-T1+T2                                    0.3534     0.2105   1.679
   0.0932 .

## Condition-L2L1+L2Only                         0.8108     0.3797   2.135
   0.0327 *

## Exposures                                    -2.8766     1.1800  -2.438
   0.0148 *

## Time-T1+T2:Condition-L2L1+L2Only             -0.1249     0.3576  -0.349
   0.7270

## Time-T1+T2:Exposures                         -0.5294     1.1645  -0.455
   0.6494

## Condition-L2L1+L2Only:Exposures               0.6818     1.7773   0.384
   0.7012

## Time-T1+T2:Condition-L2L1+L2Only:Exposures  -1.6325     2.3050  -0.708
   0.4788

## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## Correlation of Fixed Effects:

##                      (Intr) Tm-T1+T2 Cn-L2L1+L2O Expsrs Tm-T1+T2:C-L2L1
+L2O T-T1+T2:E C-L2L1+L2O:

## Time-T1+T2            0.067

## Cn-L2L1+L2O           0.236 -0.035

## Exposures            -0.207  0.042   -0.466

## Tm-T1+T2:C-L2L1+L2O   0.048  0.292    0.235      -0.095

## Tm-T1+T2:Ex          -0.087 -0.256   -0.099      -0.106 -0.595

## C-L2L1+L2O:          -0.288 -0.104   -0.087       0.253  0.031
     -0.173

## T-T1+T2:C-L2L1+L2O: -0.092 -0.515    0.061      -0.286 -0.220
      0.442    -0.012
```

```
confint <- confint(model, method = "Wald")

confint [47:54, 1:2]
```

```
##                                                    2.5 %     97.5 %

## (Intercept)                                   0.74556809  1.9563685

## Time-T1+T2                                   -0.05923436  0.7660638

## Condition-L2L1+L2Only                         0.06658643  1.5550594

## Exposures                                    -5.18944726 -0.5637463

## Time-T1+T2:Condition-L2L1+L2Only             -0.82568840  0.5759742

## Time-T1+T2:Exposures                         -2.81176739  1.7529678

## Condition-L2L1+L2Only:Exposures              -2.80151210  4.1652091

## Time-T1+T2:Condition-L2L1+L2Only:Exposures -6.15017402  2.8852507
```

Examining possible moderation effects of Dutch and Turkish vocabulary skills

In this model, Dutch and Turkish vocabulary scores, as assessed with the Diagnostic Test of Bilingualism, were added as covariates to explore possible moderation effects on the above-

reported main effect of condition. Scores on the target word retention (0 or 1) were entered as the dependent variable, condition (L2-only or L2-L1) and time (post-test 1 or post-test 2) as within-participants fixed effects, and the number of exposures, Dutch vocabulary scores, and a Turkish vocabulary score as fixed controlling factors. Condition, time and number of exposures, as well as all their possible interactions, were included as random slopes for participant, because they were within-participant fixed effects. Condition, time and number of exposures, but not their possible interactions, were included as random slopes for item, because they were within-participant fixed effects as well.

```
model <- glmer(Score ~ Time * Condition * DutchV * TurkishV * Exposures +
 (Exposures * Time * Condition | Subject) + (Exposures + Time + Condition
| Item), data=table, family="binomial", REML=FALSE, glmerControl(optimize
r="bobyqa", optCtrl = list(maxfun = 100000)))

summary (model)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace Appr
oximation) ['glmerMod']

##  Family: binomial  ( logit )

## Formula: Score ~ Time * Condition * DutchV * TurkishV * Exposures + (E
xposures *      Time * Condition | Subject) + (Exposures + Time + Conditi
on |

##     Item)

##    Data: table

## Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e
+05))

##

##      AIC      BIC   logLik deviance df.resid

##   2277.8   2716.2  -1060.9   2121.8     1960

##

## Scaled residuals:

##     Min      1Q  Median      3Q     Max

## -8.5666 -0.6609  0.2589  0.5412  2.8430

##

## Random effects:

##  Groups  Name                                        Variance Std.Dev.
Corr

##  Subject (Intercept)                                  0.45677 0.6758


##          Exposures                                   17.80752 4.2199
-0.55

##          Time-T1+T2                                   0.13916 0.3730
 0.01  0.60
```

```
##           Condition-L2L1+L2Only                          2.03204 1.4255
 0.60 -0.57  0.31

##           Exposures:Time-T1+T2                           4.56769 2.1372
-0.71 -0.07 -0.71 -0.58

##           Exposures:Condition-L2L1+L2Only                7.34690 2.7105
 0.78 -0.10  0.63  0.71 -0.98

##           Time-T1+T2:Condition-L2L1+L2Only               0.45398 0.6738
 0.41 -0.26  0.60  0.94 -0.66  0.74

##           Exposures:Time-T1+T2:Condition-L2L1+L2Only 36.02558 6.0021
-0.10 -0.77 -0.70  0.27  0.60 -0.45  0.05

##  Item     (Intercept)                                    1.14262 1.0689


##           Exposures                                      1.05499 1.0271
-0.14

##           Time-T1+T2                                     0.05308 0.2304
-0.40 -0.86

##           Condition-L2L1+L2Only                          0.26677 0.5165
 0.19  0.95 -0.98

## Number of obs: 2038, groups:  Subject, 67; Item, 19

##

## Fixed effects:

##                                                             Estimate S
td. Error z value Pr(>|z|)

## (Intercept)                                                 1.183330
 0.297541   3.977 6.98e-05 ***

## Time-T1+T2                                                  0.170260
 0.229439   0.742  0.45805

## Condition-L2L1+L2Only                                       0.931408
 0.354156   2.630  0.00854 **

## DutchV                                                      0.649146
 0.290322   2.236  0.02535 *

## TurkishV                                                    0.790555
 0.252942   3.125  0.00178 **

## Exposures                                                  -3.112260
 1.157585  -2.689  0.00718 **

## Time-T1+T2:Condition-L2L1+L2Only                            0.041592
 0.402093   0.103  0.91761

## Time-T1+T2:DutchV                                          -0.609518
 0.354055  -1.722  0.08515 .

## Condition-L2L1+L2Only:DutchV                                0.194066
 0.562418   0.345  0.73005
```

```
## Time-T1+T2:TurkishV                                          0.231703
 0.329362   0.703  0.48175

## Condition-L2L1+L2Only:TurkishV                              -0.179204
 0.514339  -0.348  0.72753

## DutchV:TurkishV                                              0.938309
 0.480720   1.952  0.05095 .

## Time-T1+T2:Exposures                                        -0.727128
 1.275197  -0.570  0.56854

## Condition-L2L1+L2Only:Exposures                              2.544099
 1.796783   1.416  0.15680

## DutchV:Exposures                                            -2.401346
 1.803645  -1.331  0.18306

## TurkishV:Exposures                                           3.532534
 1.878231   1.881  0.06000 .

## Time-T1+T2:Condition-L2L1+L2Only:DutchV                      0.066375
 0.701451   0.095  0.92461

## Time-T1+T2:Condition-L2L1+L2Only:TurkishV                    0.006928
 0.646396   0.011  0.99145

## Time-T1+T2:DutchV:TurkishV                                   0.759802
 0.591289   1.285  0.19880

## Condition-L2L1+L2Only:DutchV:TurkishV                        0.372333
 0.914391   0.407  0.68387

## Time-T1+T2:Condition-L2L1+L2Only:Exposures                   0.477660
 2.655419   0.180  0.85725

## Time-T1+T2:DutchV:Exposures                                 -0.348731
 2.160841  -0.161  0.87179

## Condition-L2L1+L2Only:DutchV:Exposures                       4.878317
 3.097034   1.575  0.11522

## Time-T1+T2:TurkishV:Exposures                               -1.804366
 2.360322  -0.764  0.44459

## Condition-L2L1+L2Only:TurkishV:Exposures                    -6.593783
 3.299636  -1.998  0.04568 *

## DutchV:TurkishV:Exposures                                    4.092528
 3.485132   1.174  0.24028

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:TurkishV             -0.739718
 1.166898  -0.634  0.52613

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:Exposures            5.550967
 4.416076   1.257  0.20876

## Time-T1+T2:Condition-L2L1+L2Only:TurkishV:Exposures          -6.111833
 4.806696  -1.272  0.20354

## Time-T1+T2:DutchV:TurkishV:Exposures                         0.909497
 4.435841   0.205  0.83755
```

```
## Condition-L2L1+L2Only:DutchV:TurkishV:Exposures            -7.056626
 6.807824  -1.037  0.29995

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:TurkishV:Exposures -4.295353
 9.192238  -0.467  0.64030

## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## Correlation matrix not shown by default, as p = 32 > 12.

## Use print(x, correlation=TRUE)  or

##   vcov(x)     if you need it
```

```
confint <- confint(model, method = "Wald")

confint [47:78, 1:2]
```

```
##                                                                2.5
 %      97.5 %

## (Intercept)                                                0.6001599
90  1.76649960

## Time-T1+T2                                                -0.2794324
52  0.61995171

## Condition-L2L1+L2Only                                      0.2372739
04  1.62554166

## DutchV                                                     0.0801255
05  1.21816599

## TurkishV                                                   0.2947974
06  1.28631352

## Exposures                                                 -5.3810843
56 -0.84343475

## Time-T1+T2:Condition-L2L1+L2Only                          -0.7464958
14  0.82968063

## Time-T1+T2:DutchV                                         -1.3034529
74  0.08441687

## Condition-L2L1+L2Only:DutchV                              -0.9082527
78  1.29638524

## Time-T1+T2:TurkishV                                       -0.4138343
04  0.87724100

## Condition-L2L1+L2Only:TurkishV                            -1.1872899
50  0.82888243

## DutchV:TurkishV                                           -0.0038841
02  1.88050272

## Time-T1+T2:Exposures                                      -3.2264676
57  1.77221141
```

```
## Condition-L2L1+L2Only:Exposures                                  -0.9775314
08  6.06572921

## DutchV:Exposures                                                 -5.9364258
91  1.13373304

## TurkishV:Exposures                                               -0.1487317
00  7.21379938

## Time-T1+T2:Condition-L2L1+L2Only:DutchV                          -1.3084443
55  1.44119341

## Time-T1+T2:Condition-L2L1+L2Only:TurkishV                        -1.2599849
27  1.27384078

## Time-T1+T2:DutchV:TurkishV                                       -0.3991042
42  1.91870774

## Condition-L2L1+L2Only:DutchV:TurkishV                            -1.4198409
46  2.16450625

## Time-T1+T2:Condition-L2L1+L2Only:Exposures                       -4.7268654
94  5.68218586

## Time-T1+T2:DutchV:Exposures                                      -4.5839022
55  3.88643931

## Condition-L2L1+L2Only:DutchV:Exposures                           -1.1917569
28 10.94839134

## Time-T1+T2:TurkishV:Exposures                                    -6.4305118
17  2.82178075

## Condition-L2L1+L2Only:TurkishV:Exposures                        -13.0609501
74 -0.12661623

## DutchV:TurkishV:Exposures                                        -2.7382058
50 10.92326193

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:TurkishV                 -3.0267958
13  1.54736080

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:Exposures                -3.1043833
73 14.20631670

## Time-T1+T2:Condition-L2L1+L2Only:TurkishV:Exposures              -15.5327833
46  3.30911779

## Time-T1+T2:DutchV:TurkishV:Exposures                             -7.7845919
28  9.60358588

## Condition-L2L1+L2Only:DutchV:TurkishV:Exposures                  -20.3997165
66  6.28646476

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:TurkishV:Exposures       -22.3118070
39 13.72110201
```

Children's enjoyment and robot preferences

To investigate whether children's robot preference affected learning gains differentially between the conditions, preference was added as a between-participant factor in this third generalized linear regression model. As before, this model took scores from the target word retention (0 or 1) as a dependent variable, condition (L2-only or L2-L1) and time (post-test 1 or post-test 2) as within-participants fixed effects, robot preference (preference for monolingual or preference for bilingual) as a between-participants fixed effect, and the number of exposures, a Dutch vocabulary score and a Turkish vocabulary score as fixed controlling factors. Condition, time and number of exposures, but not their possible interactions, were included as random slopes for participant, because they were within-participant fixed effects. Only the number of exposures was included as a random slope for item, as it was a within-participant fixed effect.

```
model <- glmer(Score ~ Time * Condition * DutchV * TurkishV * Pref * Expo
sures + (Exposures + Time + Condition | Subject) + (Exposures| Item), dat
a=table, family="binomial", REML=FALSE, glmerControl(optimizer="bobyqa",
optCtrl = list(maxfun = 100000)))

summary (model)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace Appr
oximation) ['glmerMod']

##  Family: binomial  ( logit )

## Formula: Score ~ Time * Condition * DutchV * TurkishV * Pref * Exposur
es +      (Exposures + Time + Condition | Subject) + (Exposures | Item)

##    Data: table

## Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e
+05))

##

##      AIC      BIC   logLik deviance df.resid

##   2246.0   2678.7  -1046.0   2092.0     1961

##

## Scaled residuals:

##     Min      1Q  Median      3Q     Max

## -5.8872 -0.6728  0.2624  0.5520  4.3222

##

## Random effects:

##  Groups  Name                   Variance Std.Dev. Corr

##  Subject (Intercept)            0.55270  0.7434

##          Exposures              6.08200  2.4662   -1.00

##          Time-T1+T2             0.03963  0.1991    0.05 -0.02

##          Condition-L2L1+L2Only 1.59744  1.2639    0.76 -0.74  0.68

##  Item    (Intercept)            1.24755  1.1169

##          Exposures              2.86081  1.6914   0.07
```

```
## Number of obs: 2038, groups:  Subject, 67; Item, 19

##

## Fixed effects:

##
    Estimate Std. Error z value Pr(>|z|)

## (Intercept)
    1.358877   0.309315   4.393 1.12e-05 ***

## Time-T1+T2
    0.198853   0.204119   0.974 0.329957

## Condition-L2L1+L2Only
    0.993753   0.298341   3.331 0.000866 ***

## DutchV
    0.915802   0.300828   3.044 0.002332 **

## TurkishV
    0.861309   0.272242   3.164 0.001557 **

## Pref-Bi+Mono
    0.423046   0.308815   1.370 0.170718

## Exposures
   -4.617483   1.178855  -3.917 8.97e-05 ***

## Time-T1+T2:Condition-L2L1+L2Only
    0.005065   0.394389   0.013 0.989753

## Time-T1+T2:DutchV
   -0.849786   0.391623  -2.170 0.030014 *

## Condition-L2L1+L2Only:DutchV
   -0.330542   0.549808  -0.601 0.547709

## Time-T1+T2:TurkishV
   -0.060455   0.340447  -0.178 0.859056

## Condition-L2L1+L2Only:TurkishV
   -0.272974   0.497350  -0.549 0.583103

## DutchV:TurkishV
    0.866933   0.499156   1.737 0.082423 .

## Time-T1+T2:Pref-Bi+Mono
    0.210078   0.398776   0.527 0.598328

## Condition-L2L1+L2Only:Pref-Bi+Mono
    0.253458   0.578701   0.438 0.661403

## DutchV:Pref-Bi+Mono
    0.650049   0.591757   1.099 0.271983

## TurkishV:Pref-Bi+Mono
   -0.312439   0.538636  -0.580 0.561877

## Time-T1+T2:Exposures
    1.278651   1.719664   0.744 0.457151
```

```
## Condition-L2L1+L2Only:Exposures
   -1.375902   1.937566  -0.710 0.477631

## DutchV:Exposures
    0.359838   1.703527   0.211 0.832707

## TurkishV:Exposures
    1.246110   1.728138   0.721 0.470866

## Pref-Bi+Mono:Exposures
   -0.986282   2.047001  -0.482 0.629935

## Time-T1+T2:Condition-L2L1+L2Only:DutchV
    0.730414   0.770329   0.948 0.343035

## Time-T1+T2:Condition-L2L1+L2Only:TurkishV
   -0.452818   0.681995  -0.664 0.506716

## Time-T1+T2:DutchV:TurkishV
    1.137462   0.656129   1.734 0.082990 .

## Condition-L2L1+L2Only:DutchV:TurkishV
   -0.065529   0.938913  -0.070 0.944359

## Time-T1+T2:Condition-L2L1+L2Only:Pref-Bi+Mono
    0.784138   0.794814   0.987 0.323855

## Time-T1+T2:DutchV:Pref-Bi+Mono
   -0.463393   0.777956  -0.596 0.551406

## Condition-L2L1+L2Only:DutchV:Pref-Bi+Mono
    0.836192   1.112277   0.752 0.452181

## Time-T1+T2:TurkishV:Pref-Bi+Mono
   -0.376606   0.694679  -0.542 0.587729

## Condition-L2L1+L2Only:TurkishV:Pref-Bi+Mono
   -0.280227   0.989432  -0.283 0.777008

## DutchV:TurkishV:Pref-Bi+Mono
   -3.220065   0.995842  -3.234 0.001223 **

## Time-T1+T2:Condition-L2L1+L2Only:Exposures
    5.043025   3.417431   1.476 0.140031

## Time-T1+T2:DutchV:Exposures
   -3.193825   2.721708  -1.173 0.240610

## Condition-L2L1+L2Only:DutchV:Exposures
    1.707101   3.148187   0.542 0.587647

## Time-T1+T2:TurkishV:Exposures
   -1.079184   2.825102  -0.382 0.702463

## Condition-L2L1+L2Only:TurkishV:Exposures
  -11.135966   3.284772  -3.390 0.000698 ***

## DutchV:TurkishV:Exposures
    8.357200   4.411413   1.894 0.058165 .

## Time-T1+T2:Pref-Bi+Mono:Exposures
    1.118419   3.463063   0.323 0.746728
```

```
## Condition-L2L1+L2Only:Pref-Bi+Mono:Exposures
   -2.040798   3.799390  -0.537 0.591172

## DutchV:Pref-Bi+Mono:Exposures
    3.708957   3.433646   1.080 0.280062

## TurkishV:Pref-Bi+Mono:Exposures
   -5.515676   3.381281  -1.631 0.102840

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:TurkishV
   -0.881224   1.321526  -0.667 0.504885

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:Pref-Bi+Mono
    2.498896   1.543580   1.619 0.105470

## Time-T1+T2:Condition-L2L1+L2Only:TurkishV:Pref-Bi+Mono
   -2.243219   1.364405  -1.644 0.100155

## Time-T1+T2:DutchV:TurkishV:Pref-Bi+Mono
    2.004787   1.375159   1.458 0.144880

## Condition-L2L1+L2Only:DutchV:TurkishV:Pref-Bi+Mono
   -0.536728   1.880964  -0.285 0.775378

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:Exposures
    5.256591   5.511036   0.954 0.340170

## Time-T1+T2:Condition-L2L1+L2Only:TurkishV:Exposures
    2.702789   5.607784   0.482 0.629827

## Time-T1+T2:DutchV:TurkishV:Exposures
    4.796830   7.748565   0.619 0.535877

## Condition-L2L1+L2Only:DutchV:TurkishV:Exposures
    3.862213   8.551349   0.452 0.651521

## Time-T1+T2:Condition-L2L1+L2Only:Pref-Bi+Mono:Exposures
    7.336599   6.861534   1.069 0.284963

## Time-T1+T2:DutchV:Pref-Bi+Mono:Exposures
   -8.968633   5.409439  -1.658 0.097326 .

## Condition-L2L1+L2Only:DutchV:Pref-Bi+Mono:Exposures
    4.893231   6.343906   0.771 0.440513

## Time-T1+T2:TurkishV:Pref-Bi+Mono:Exposures
    3.536507   5.537586   0.639 0.523059

## Condition-L2L1+L2Only:TurkishV:Pref-Bi+Mono:Exposures
   -3.790778   6.586364  -0.576 0.564920

## DutchV:TurkishV:Pref-Bi+Mono:Exposures
    7.543731   9.114680   0.828 0.407871

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:TurkishV:Pref-Bi+Mono
   -0.275670   2.733595  -0.101 0.919673

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:TurkishV:Exposures
  -13.977319  14.983240  -0.933 0.350890

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:Pref-Bi+Mono:Exposures
   -2.874791  11.017727  -0.261 0.794151
```

```
## Time-T1+T2:Condition-L2L1+L2Only:TurkishV:Pref-Bi+Mono:Exposures
   15.073230  11.260020   1.339 0.180685

## Time-T1+T2:DutchV:TurkishV:Pref-Bi+Mono:Exposures
    7.604616  15.836608   0.480 0.631091

## Condition-L2L1+L2Only:DutchV:TurkishV:Pref-Bi+Mono:Exposures
   52.163526  16.983393   3.071 0.002130 **

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:TurkishV:Pref-Bi+Mono:Exposure
s -29.470130  31.676647  -0.930 0.352194

## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## Correlation matrix not shown by default, as p = 64 > 12.

## Use print(x, correlation=TRUE)  or

##   vcov(x)     if you need it

## convergence code: 1

## Model failed to converge with max|grad| = 0.00375898 (tol = 0.001, com
ponent 1)
```

```
confint <- confint(model, method = "Wald")

confint [14:77, 1:2]
```

```
##
        2.5 %      97.5 %

## (Intercept)
    0.7526312  1.96512180

## Time-T1+T2
   -0.2012136  0.59892013

## Condition-L2L1+L2Only
    0.4090151  1.57848996

## DutchV
    0.3261897  1.50541467

## TurkishV
    0.3277255  1.39489279

## Pref-Bi+Mono
   -0.1822202  1.02831270

## Exposures
   -6.9279971 -2.30696912

## Time-T1+T2:Condition-L2L1+L2Only
   -0.7679231  0.77805313

## Time-T1+T2:DutchV
   -1.6173541 -0.08221857

## Condition-L2L1+L2Only:DutchV
   -1.4081459  0.74706104
```

```
## Time-T1+T2:TurkishV
   -0.7277182  0.60680809

## Condition-L2L1+L2Only:TurkishV
   -1.2477618  0.70181305

## DutchV:TurkishV
   -0.1113953  1.84526141

## Time-T1+T2:Pref-Bi+Mono
   -0.5715088  0.99166498

## Condition-L2L1+L2Only:Pref-Bi+Mono
   -0.8807752  1.38769074

## DutchV:Pref-Bi+Mono
   -0.5097724  1.80987057

## TurkishV:Pref-Bi+Mono
   -1.3681470  0.74326830

## Time-T1+T2:Exposures
   -2.0918286  4.64913128

## Condition-L2L1+L2Only:Exposures
   -5.1734615  2.42165772

## DutchV:Exposures
   -2.9790129  3.69868855

## TurkishV:Exposures
   -2.1409792  4.63319873

## Pref-Bi+Mono:Exposures
   -4.9983306  3.02576661

## Time-T1+T2:Condition-L2L1+L2Only:DutchV
   -0.7794030  2.24023181

## Time-T1+T2:Condition-L2L1+L2Only:TurkishV
   -1.7895046  0.88386865

## Time-T1+T2:DutchV:TurkishV
   -0.1485285  2.42345171

## Condition-L2L1+L2Only:DutchV:TurkishV
   -1.9057657  1.77470757

## Time-T1+T2:Condition-L2L1+L2Only:Pref-Bi+Mono
   -0.7736690  2.34194508

## Time-T1+T2:DutchV:Pref-Bi+Mono
   -1.9881586  1.06137191

## Condition-L2L1+L2Only:DutchV:Pref-Bi+Mono
   -1.3438312  3.01621614

## Time-T1+T2:TurkishV:Pref-Bi+Mono
   -1.7381514  0.98493900

## Condition-L2L1+L2Only:TurkishV:Pref-Bi+Mono
   -2.2194790  1.65902512
```

```
## DutchV:TurkishV:Pref-Bi+Mono
   -5.1718795 -1.26825097

## Time-T1+T2:Condition-L2L1+L2Only:Exposures
   -1.6550175 11.74106648

## Time-T1+T2:DutchV:Exposures
   -8.5282754  2.14062561

## Condition-L2L1+L2Only:DutchV:Exposures
   -4.4632332  7.87743471

## Time-T1+T2:TurkishV:Exposures
   -6.6162822  4.45791364

## Condition-L2L1+L2Only:TurkishV:Exposures
  -17.5740006 -4.69793113

## DutchV:TurkishV:Exposures
   -0.2890104 17.00341100

## Time-T1+T2:Pref-Bi+Mono:Exposures
   -5.6690609  7.90589813

## Condition-L2L1+L2Only:Pref-Bi+Mono:Exposures
   -9.4874667  5.40587015

## DutchV:Pref-Bi+Mono:Exposures
   -3.0208653 10.43877874

## TurkishV:Pref-Bi+Mono:Exposures
  -12.1428640  1.11151230

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:TurkishV
   -3.4713676  1.70891881

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:Pref-Bi+Mono
   -0.5264654  5.52425722

## Time-T1+T2:Condition-L2L1+L2Only:TurkishV:Pref-Bi+Mono
   -4.9174030  0.43096594

## Time-T1+T2:DutchV:TurkishV:Pref-Bi+Mono
   -0.6904746  4.70004803

## Condition-L2L1+L2Only:DutchV:TurkishV:Pref-Bi+Mono
   -4.2233486  3.14989297

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:Exposures
   -5.5448416 16.05802327

## Time-T1+T2:Condition-L2L1+L2Only:TurkishV:Exposures
   -8.2882655 13.69384305

## Time-T1+T2:DutchV:TurkishV:Exposures
  -10.3900789 19.98373947

## Condition-L2L1+L2Only:DutchV:TurkishV:Exposures
  -12.8981223 20.62254822

## Time-T1+T2:Condition-L2L1+L2Only:Pref-Bi+Mono:Exposures
   -6.1117592 20.78495795
```

```
## Time-T1+T2:DutchV:Pref-Bi+Mono:Exposures
  -19.5709389   1.63367201

## Condition-L2L1+L2Only:DutchV:Pref-Bi+Mono:Exposures
   -7.5405954 17.32705781

## Time-T1+T2:TurkishV:Pref-Bi+Mono:Exposures
   -7.3169626 14.38997718

## Condition-L2L1+L2Only:TurkishV:Pref-Bi+Mono:Exposures
  -16.6998145   9.11825817

## DutchV:TurkishV:Pref-Bi+Mono:Exposures
  -10.3207135 25.40817462

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:TurkishV:Pref-Bi+Mono
   -5.6334175   5.08207818

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:TurkishV:Exposures
  -43.3439299 15.38929242

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:Pref-Bi+Mono:Exposures
  -24.4691384 18.71955698

## Time-T1+T2:Condition-L2L1+L2Only:TurkishV:Pref-Bi+Mono:Exposures
   -6.9960038 37.14246330

## Time-T1+T2:DutchV:TurkishV:Pref-Bi+Mono:Exposures
  -23.4345644 38.64379635

## Condition-L2L1+L2Only:DutchV:TurkishV:Pref-Bi+Mono:Exposures
   18.8766867 85.45036488

## Time-T1+T2:Condition-L2L1+L2Only:DutchV:TurkishV:Pref-Bi+Mono:Exposure
s -91.5552176 32.61495797
```