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

D6.2: Output module for space domain

Due Date: **31/03/2018**
Submission Date: **02/07/2018**

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

Organisation name of lead contractor for this deliverable: **Tilburg University**

Responsible Person: **Emiel Krahmer** 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

This deliverable describes the output module for the space domain and general changes to the module since Deliverable 6.1. We describe how the revised objectives have come to affect the development of the output module for the large-scale evaluation study. We explain how various (rudimentary) functionalities have been developed further and finalized for the purpose of this evaluation. We discuss the ways in which we have circumvented several limiting factors of the current setup. Finally, we argue for more exploration of data-driven approaches to the generation of robot behaviours – so as to avoid a designers' bias and reduce workloads – and briefly mention related work in progress.

Principal Contributors

The main authors of this deliverable are as follows:

Jan de Wit, Tilburg University
Bram Willemsen, Tilburg University
Mirjam de Haas, Tilburg University
Emiel Krahmer, Tilburg University
Paul Vogt, Tilburg University

Thorsten Schodde, Bielefeld University

Aylin C. Küntay, Koç University
Tilbe Göksun, Koç University
Özlem Ece Demir-Lira, Koç University
Junko Kanero, Koç University

Revision History

Version 1.0 (02-07-2018)
First version.

1 Introduction

Work Package 6 focusses on developing an output module, which is the part of the L2TOR tutoring system concerned with the verbal and non-verbal communicative behaviour of the robot. As a humanoid agent, the robot has the ability to interact with people using spoken natural language. The content of the robot's utterances will need to be generated to best suit the purpose, in case of the L2TOR project to provide the means, and a context, for children to learn a second language. There are several options for implementing natural language generation (NLG) [1]. For the L2TOR system we implemented template-based language generation, where a mostly static template of an utterance is complemented with task-related information, such as the name of the conversational partner or specific objects that are involved in the completion of a task [2, 3]. The choice for a template-based solution was made because alternative options (e.g., a machine learning approach) tend to result in output that is less constrained and, therefore, less predictable. The use of templates ensures that learners are exposed to similar content, which allows for a fair comparison between their experiences with the system and their learning outcomes. Furthermore, because the aim of the tutoring system is to teach a language to a young target audience, it was deemed important to avoid inappropriate or (grammatically) incorrect language at all costs.

Non-verbal behaviours were designed and implemented as well. This includes the use of the robot's gaze to direct the attention of the child, and to help build rapport [4, 5]. Furthermore, the robot is able to perform actions on the tablet, which are accompanied by a gesture to indicate the robot's intention (e.g., pretending to select an object by "tapping" on the screen). Because we noticed in a previous experiment [6] (see Appendix) that the robot's use of iconic gestures helped children with long-term memorization of new words, a condition where the robot would perform iconic gestures was also included in the large-scale study (further detailed in WP7).

The output module was largely functional at the time Deliverable 6.1 was submitted, however it was only briefly tested in a pilot setting with only one lesson implemented. Since then, the system has seen several more iterations so that it would support all of the planned lesson series and fulfil the technical requirements for the large-scale study. This includes features that were (partially) left unfinished in the previous version, new additions to the module, and future directions that were listed in Deliverable 6.1. The integrated tutoring system has recently been evaluated in several (pilot) studies and it is currently being used in the large-scale study. The details of this large-scale study have been described in the *Revised objectives* document that was submitted on November 3rd, 2017.

While developing the various modules of the integrated tutoring system, we discovered that virtually all of the planned functionalities were not specifically designed for one particular lesson or domain. Instead, the technical implementation generalizes to a broad range of lesson content because most of the lessons rely on the same functionality and interaction patterns (e.g., manipulating objects on the tablet and repeating L2 utterances). By using abstractions such as "tasks" and "objectives", it is possible to create a platform that can be easily extended to support new content. Therefore, instead of focussing on work that was done on the space domain in the context of Work Package 6, we talk about the overall progress of technical developments within WP 6, as this will automatically cover what has changed within the system to accommodate the lessons of the space domain.

The following sections will recapitulate the architecture of the integrated tutoring system as it is being deployed in the large-scale study, focusing on additions and changes to this system since D6.1. Furthermore, several studies that are currently being planned and executed within the context of WP6 will be highlighted.

2 Revised System Architecture

The overall system architecture, and the position of the OutputManager, remains as described in D6.1:

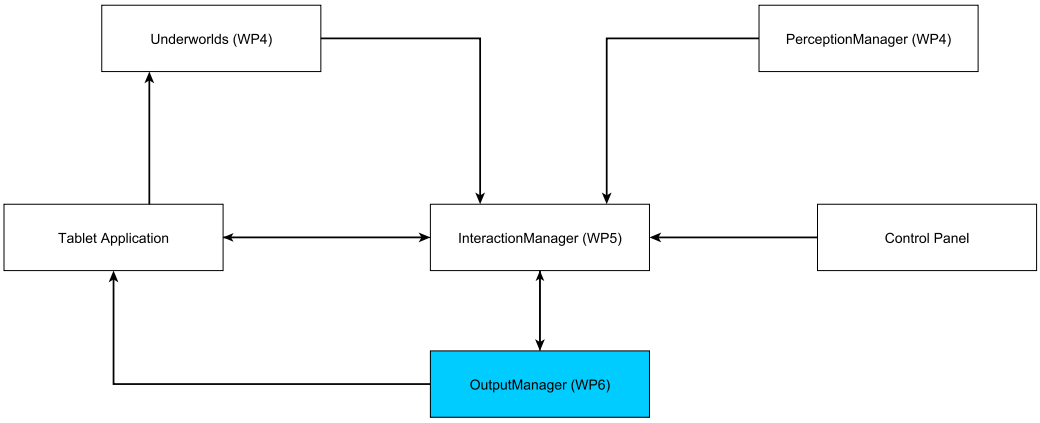


Figure 1: The integrated technical tutoring system and the role of the OutputManager.

It should be noted once more that there is a ConnectionManager, developed within WP3, working behind the scenes to facilitate the communication between the various modules. Although the overall architecture did not change, the features that were added to the OutputManager (as described in the following sections) result in an increase in the amount of messages being sent between the OutputManager and the InteractionManager and Tablet Application. However, the modules remain independent, so that changes in one module do not directly affect all other modules, as long as communication between modules occurs in the format that was agreed upon. This also allows modules to be replaced by other versions, for example the OutputManager should not be affected if the current InteractionManager was to be replaced by another version that adaptively goes through the storyboard rather than linearly, or if we were to exchange the tablet game with another platform. This is possible because each module generates certain outputs and expects certain inputs, as long as these are accounted for the modules will be able to communicate with each other.

2.1 Output Module

The internal architecture of the OutputManager is shown in Figure 2. The existing submodules have expanded with the new features and improvements described in the next section. One new submodule was added to set up the tablet condition, which takes care of rerouting the audio from the robot to the speakers in the tablet, and that all non-verbal behaviours from the robot are disabled. Furthermore, an iconic gesture condition was implemented inside the OutputRealizer, and the FeedbackManager now dynamically generates feedback based on the context of the current task.

3 Updates to and extensions of the Output Module

Several parts of the Output Module still had to be completed or updated before the start of the large-scale study. A shared spreadsheet was used to keep track of features that still needed to be implemented, and bugs that occurred while testing the system during various pilot studies. This allowed everyone in the technical development team, as well as colleagues running the pilot studies, to be able to see and

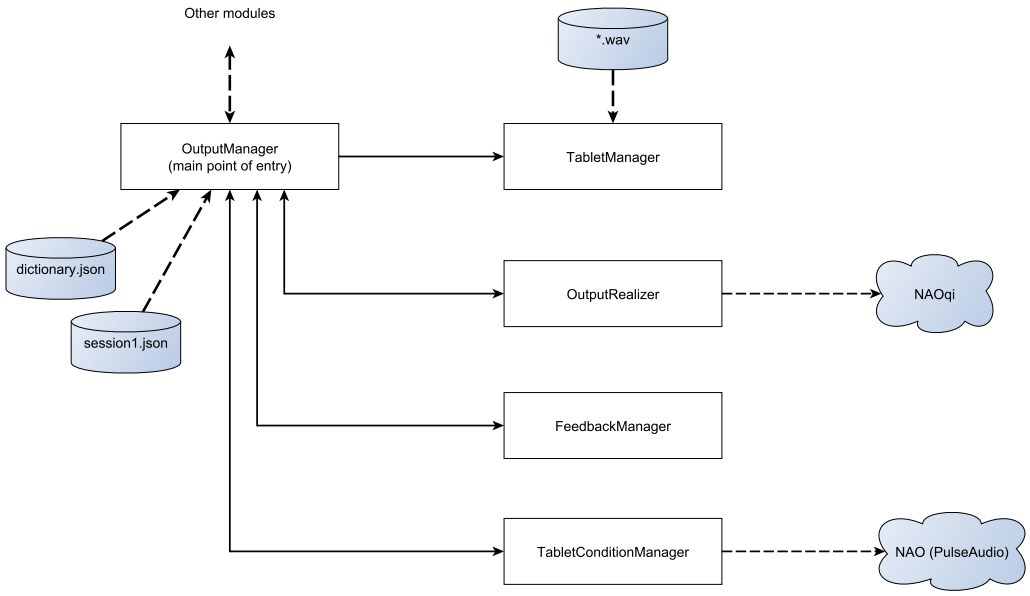


Figure 2: The internal architecture of the Output Module.

update the state of the Output Module. Priorities were assigned to each point on the list. The most prominent changes since D6.1 are discussed in detail below.

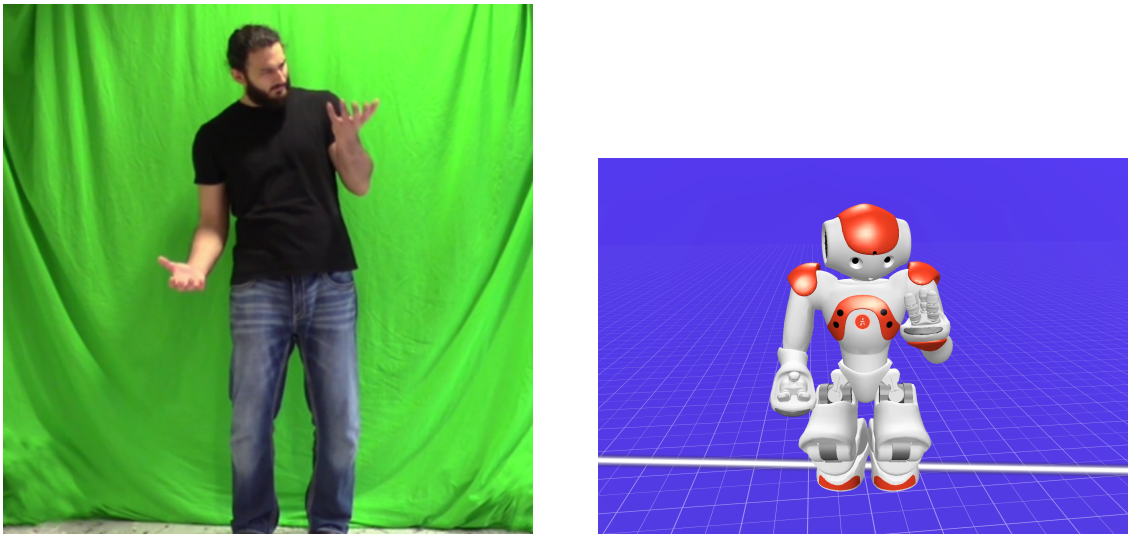
3.1 Gesture condition

In the context of WP6, we have conducted a study on the effects of (iconic) gestures on L2 vocabulary acquisition. This study, which was previously described and for which preliminary results were presented in D6.1, has since been presented at the 13th Annual ACM/IEEE International Conference on Human-Robot Interaction [6] (see Appendix).

One of our main findings was that children who had been presented with novel L2 words in combination with iconic gestures were, on average, able to retain their newly acquired knowledge to a higher degree than children who had not been exposed to these iconic gestures. This reaffirmed the notion that the use of gestures can indeed be an effective scaffolding technique (e.g, [7, 8, 9, 10, 11]), even when learning from a robot rather than a human tutor.

However, our study [6] was limited in the sense that the effects were only measured for a single tutoring session. Whether these positive effects hold true when children interact with a robot tutor over a longer period of time (i.e., multiple sessions) and for L2 words of various other domains is unclear. Therefore, the decision was made to address these questions during the large-scale evaluation study by introducing an additional experimental condition that would let us examine in isolation the effects of (iconic) gesture use versus the mere presence of a more static robot over the course of multiple tutoring sessions and for L2 words of various domains. This means that four experimental conditions were defined for the large-scale evaluation study, namely a control condition, a tablet-only condition, a robot without (iconic) gestures condition, and a robot with (iconic) gestures condition.

The decision to include iconic gestures meant that for each of the 34 target words, a gesture that was acceptably iconic had to be designed and implemented, to be presented with the appropriate L2 target word during the tutoring interactions. To this end, students of Koc University were recorded while



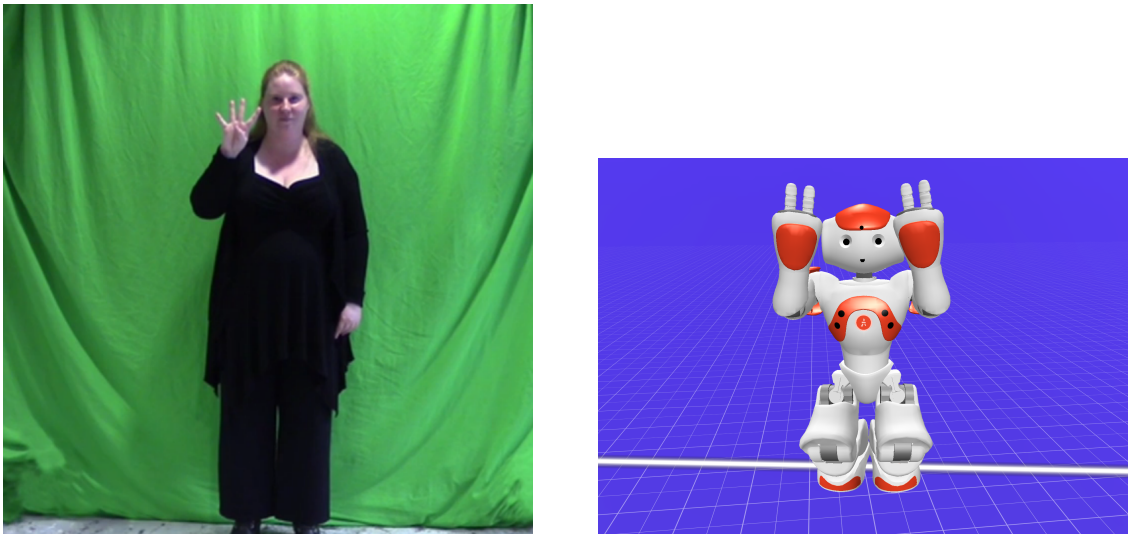
(a) *light*, as performed by a human

(b) *light*, robot implementation

Figure 3: Example of gesture mapping from human to robot for the concept *light*. On the left, a gesture for *light*, as performed by a human; on the right, the implementation of the depicted human gesture.

performing gestures for the various target words. As some of the target words concern concepts that are quite difficult to express non-verbally, the resulting gestures had varying degrees of iconicity. An additional constraint was the mapping of these gestures, by means of puppeteering (this implementation technique was previously explained in D6.1), to the NAO robot: given its limited degrees of freedom, this proved challenging. For reference, see Figures 3 and 4, in which stills from gestures for target words *light* and *four* are shown, performed by a human as well as the eventual robot implementation (image from the virtual NAO in the Choregraphe workspace). Students of Tilburg University who had followed a course on HRI were asked to implement gestures based of the video recordings. A number of these implementations were used as the basis for gestures used in the actual large-scale evaluation study, after being further refined by members of both the Koc and Tilburg teams; others were redesigned and implemented from scratch (an example of which is shown in Figure 4, as the human-produced gesture uses four fingers on one hand to indicate the target word *four*; a gesture that is impossible for the NAO to perform as it only has three fingers on one hand).

The final challenge was to incorporate the gestures in the interaction without disrupting the flow. The system's architecture had initially not been designed to facilitate the use of gestures as required for the large-scale evaluation study, as made apparent by complications involving the combined use of verbal and non-verbal behaviours. In addition, the physical limitations can slow down the robot's execution of behaviours, which, in turn, slows down the interaction. To mitigate these issues, we have carefully timed the execution of gestures with the accompanying speech: each time a gesture is executed, a pause is inserted to synchronize the gesture with the appropriate part of the robot's speech. Although an interaction with the use of gestures will still take slightly longer than an interaction without gestures, the interaction flow has been mostly preserved as a result of the use of these predetermined pauses.



(a) *four*, as performed by a human

(b) *four*, robot implementation

Figure 4: Example of gesture mapping from human to robot for the concept *four*. On the left, a gesture for *four*, as performed by a human; on the right, the implementation of the depicted human gesture. Due to the robots physical limitations, a direct mapping was not possible.

3.2 Tablet condition

For the experimental condition where children would play only with the tablet, without the robot being physically present, a version of the system would have to be designed with as little modifications as possible so that results between the conditions with and without robot could be compared. The proposed solution was to simply have all the speech output come from the speakers of the tablet instead of the robot, and leave the game and storyboard the same as in the robot conditions. There would not be any visual representation of a virtual agent, because the goal of including this experimental condition is to compare a tablet game and an embodied agent, rather than a virtual and a physical agent. Together with our colleagues from Utrecht University, the storyboards were checked and modified where needed to ensure that all utterances worked for all conditions, with and without a robot present. Details of these changes to the storyboards are described in D2.3.

Ideally, implementing the tablet condition would simply mean replacing the OutputRealizer submodule with one that uses a Text-to-Speech engine running on the tablet instead of the NAO robot, and simply ignores all non-verbal cues. For consistency in the quality of pronunciation and comprehensibility, the idea was to have the tablet use the same voice as the robot. After consulting with SoftBank it turned out that it would be a challenge to run this speech engine on a tablet, since it was licensed specifically to run on their robotics platforms. Fortunately, the NAO robots have a feature to stream all audio output to an external device over the network. This did allow us to play exactly the same content in all conditions, with the drawback that the robot would still have to be present (but hidden from view) for us to use its Text-to-Speech engine. An alternative option would be to record samples of the robot's voice and play them back from the tablet when needed. In this case the robot would not have to be there, however this lacks the flexibility and extensibility of the current system because all of the content and possible variations (e.g., name of the child, contextual information) would have to be generated prior to running a lesson.

The PulseAudio software that is used for the tablet condition requires two parts: a server that has

access to the audio output device that is to be used – in our case the tablet – and a client that streams the audio signal to this server. The client is pre-installed on the NAO robot, it simply needs to be configured so that it knows where to find the server, and to switch audio output to the server instead of the robot's speakers. We could not find extensive documentation on how to configure and run the software reliably, however there was a member of the L2TOR project that had worked with it before, who was able to provide valuable help with the initial set-up of the system. Although PulseAudio support on Windows platforms was previously lacking, it appears to have become more stable in recent years, allowing us to run the server on the tablet. The PulseAudio server is always started when initializing the OutputManager, even if the tablet condition is not used – the robot will simply never connect to it in this case. When the experiment is then started in the tablet condition, the process of connecting to the robot and re-configuring it to switch output from the robot's speakers to the tablet has been automated. After a lesson in the tablet condition has finished, the output is switched back to the robot's speakers. We discovered a bug where the audio would stutter, mostly at the start of the session. A work-around was discovered thanks to our target domain of language learning, where language switches are common: it turns out that switching the language removes the stuttering issues, so when initializing the tablet condition we now always change language to prevent the issue from occurring during the interaction. Our findings while getting PulseAudio to work have been communicated to SoftBank to help other projects in the future.

3.3 Context-sensitive prompts and feedback

After pilot testing with place-holder feedback such as "that's not quite right", as well as requests for answer that simply repeated the last sentence of the task introduction, it became clear that these interaction mechanisms would be much more efficient and cause less frustration if they would become more intelligent by adding context-awareness. For interactions with the tablet game, the OutputManager would already receive the required action (e.g., moving an object inside of another object) and the unique identifiers of objects involved (e.g., monkey_1 and cage_1) from the InteractionManager. This, combined with a template for each task and a dictionary that maps object identifiers to their descriptions in the correct language was already enough to present basic tasks and feedback, for example:

> Let's put the *giraffe* in the cage.

The template for this type of task, to move an object, was:

> Let's put [obj_1] [rel] [obj_2]

To find out whether the two objects involved in this task, the giraffe and the cage, should be pronounced in the L1 or L2, we look at the last utterance – where the task was originally introduced. While this implementation worked for the first simple tasks that we encountered, there were a number of situations in which a sense of context was needed, for example:

- There are several giraffes, but the target is the one that is inside a cage
- The target is the area that contains *most* animals: the feeding tray or the lake

In these cases, the information that the InteractionManager needs to verify whether the task was completed does not provide enough input to formulate a task, which is needed to request an answer or to give corrective feedback. For instance, the InteractionManager will only know that the correct answer to the "cage with most animals" is cage_3, it is unaware of the contents of this cage or how to describe it with words. To add this functionality to the OutputManager, we implemented a first version

of a discourse model: a mental model of the objects in the tablet game, and their relationship to each other at any point during the interaction. Parts of this functionality was already implemented within Underworlds (WP4), but due to the strict time schedule in which the discourse model would be needed, we decided to create a basic implementation ourselves to minimise the impact on Underworlds' code base. This also allowed us to rapidly explore and validate our requirements for the discourse model.

3.3.1 Discourse model

The discourse model is integrated with the parser that converts the Excel storyboards into JSON files that can be read by the InteractionManager and OutputManager. While going through the storyboard for a lesson, at certain times the tablet game will load a new 3D scene. A single JSON file is used by the tablet game to build this entire scene, describing the background image and the details of all 3D objects that should be created. Whenever a change in 3D scene is encountered, the parser also loads this JSON file describing the 3D world and analyses the location and visibility of all objects contained within. This is the initial state of the tablet game, of which initial spatial relations are inferred and stored (e.g., objects that have walls are considered to be containers and objects that are positioned within these walls are considered to be "in" these containers). As the storyboard plays out, the state of this 3D world is likely to change: some tasks ask the child to move an object to a new location, and some events that occur during the interaction will cause objects to be moved as well, or objects that were previously hidden can become visible and vice versa. These tasks and events are all described in the storyboard, therefore if the discourse model is updated accordingly we are able to derive the location of each object, and its spatial relation to other objects, at any point during the interaction. The discourse model was developed with a focus on practicality, there are several limitations that we intend to address in the future:

- The model currently runs offline (when parsing storyboards), but could be made to run during the interaction, allowing the OutputManager to request information from it at any time – in co-operation with Underworlds (WP4), which already maintains the spatial relations between objects in the 3D scene;
- The size of 3D objects is unknown, therefore the model cannot be sure about the initial spatial relations between all objects;
- While the model is able to determine "the area with fewest things", it does not support queries such as "item of which there are fewest" – a work-around we use now is to have an invisible dummy object which the dictionary translates into "item of which there are fewest";
- Changes in the game world due to the robot performing actions are not yet updated in the discourse model.
- Ideally, objects that are contained within other objects should also be recognized as valid targets. For example, if a cage contains giraffes, clicking on those giraffes should also be correct when the cage is the target. For now, the giraffes (and the floor of the cage) have to be added separately as targets.

After creating the discourse model to keep track of the (virtual) context, the storyboard had to be updated. Previously, the identifier of the correct answer to a task had to be listed:

<giveResponseToSelectObject(cage_3)>

After implementing the discourse model, this task has been changed to:

<giveResponseToSelectObject(cage with most animals)>

This makes the storyboard easier to read and create, because the object identifiers and the state of the 3D scene do not need to be considered to be able to understand this task. While converting this storyboard into the JSON files that are used by the system, the discourse model would translate the description of the target object into the correct identifier, cage_3, for the InteractionManager, while the OutputManager gets a data structure of the target object, such as:

```
"objective": {
    "id": "cage",
    "is_plural": false,
    "rel": {
        "target": {
            "id": "animal",
            "is_plural": true
        },
        "type": "most"
    }
}
```

This data structure can then be turned into a textual description of the object, e.g., by looking up the singular word for “cage” and the plural word for “animal” in the dictionary, in the correct language, and then combining those into “the cage with most animals”. This full description of the target object can then be plugged into the template for the current task type, which in this case is to touch an object.

3.3.2 Feedback

The basic workings of the FeedbackManager have previously been discussed in D6.1. The goal of this submodule is to deliver feedback relevant to the context of the learning interaction. In practice, this means that feedback messages are tailored to the specific task carried out by the learner. To provide the FeedbackManager with the necessary context-awareness for the tailoring of these messages, information is relayed from other modules and modified for use in specific syntactic templates. In similar vein to the functionality of requesting answers, this information is parsed and used to fill the gaps.

The current implementation of the FeedbackManager handles the requests for feedback step by step. The pseudocode shown in Figure 6 describes the process of constructing a contextually-relevant feedback message for an object selection task. To be more specific, Figure 6 highlights the generation of a feedback message, in an interaction in which the L1 is Dutch, for an object selection task for which the learner has provided an incorrect answer, and for which relationship information has been found in the previously described “objective” data structure of the target object (of which an example was provided). The feedback message will consist of either two or three parts: in case the learner has provided an incorrect answer no more than once, the feedback message will be extended with a prompt requesting the learner to try again to find the correct answer; if an incorrect answer has already been provided twice, the learner will not be prompted again, but will instead be helped and guided towards the correct answer by the tutoring system (see Section 3.4). In the first part of the feedback message it is made explicit to the learner whether or not they have given a correct answer (e.g., “That

is not quite correct"); the second part consists of the syntactic template, the gaps of which are filled with task-relevant information. Finally, the conceptual message may or may not be followed by the aforementioned prompt, depending on the number of attempts by the learner up to this point in the interaction. Figure 5 shows an annotated feedback message for an object selection task.

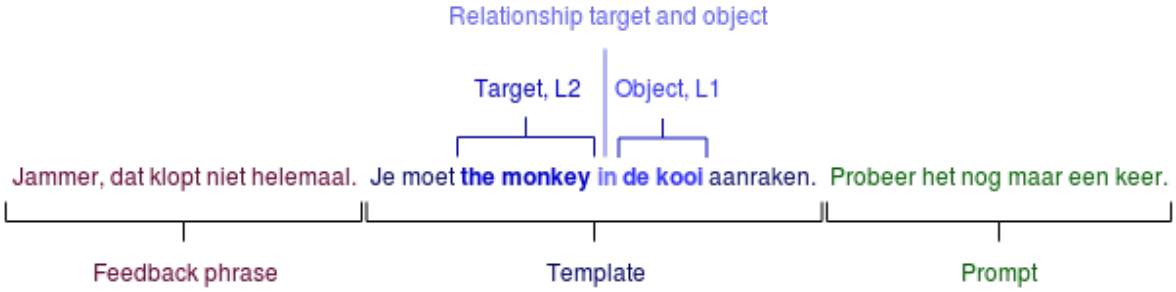


Figure 5: Example of a feedback message for an object selection task. The message consists of three parts: a (negative) feedback phrase, the syntactic template with slots filled with task-relevant information, and a prompt for the learner to attempt the task once more.

A more complete overview and motivated description of the various possible feedback messages is provided in D2.3.

```
negative feedback ⇐ list of negative feedback options
feedback count ⇐ feedback count
if object selection task then
    if L1 is Dutch then
        if answer is correct then
            ...
        else
            if relationship information in objective then
                parse objective
                filled template ⇐ syntactic template with parsed objective for slot filling
                new feedback ⇐ old feedback
                while new feedback is old feedback do
                    new feedback ⇐ random select from negative feedback
                end while
                old feedback ⇐ new feedback
                feedback message ⇐ new feedback + filled template
                if feedback count < 2 then
                    feedback message ⇐ feedback message + prompt learner
                    return feedback message
                else
                    return feedback message
                end if
            else
                ...
            end if
        end if
    else
        ...
    end if
else if ... then
    ...
end if
```

Figure 6: Highlight of procedure for the generation of a feedback message for an object selection task.

3.3.3 Use of the child's name

In several of the storyboards, the child's name was occasionally used by the robot as part of the conversation to make the interaction feel more personal, draw the attention back to the interaction and to indicate a sense of urgency:

> I think we've found all the animals! **Jane**, how many animals are there?

The storyboards were revised so that the name was mentioned more often, and the name was also included for the second and third request for answer, for example:

> **Jane**, let's put the monkey in the cage.

3.4 Providing help and manipulating the tablet

To ensure that lessons can always be completed, even if the child does not understand the current task, the robot will take over and perform the task for the child after several times of requesting an answer, or after two incorrect attempts of completing the task. This is designed to reduce frustration when a child does not succeed in a particular task. The output module receives a notification from the InteractionManager when it is time to provide help, after which the robot will start by verbally communicating its intention to give help. If a task requires action to be taken on the tablet (e.g., moving or selecting an object), the robot performs the act for the child:

> I will show you how to do it, look!

Depending on the type of interaction with the tablet, the robot will then perform a gesture that simulates dragging or tapping of an object on the screen. While these gestures take place, signals are sent to the tablet game at specific times so that it appears as though objects are actually being moved by the robot (by animating and highlighting objects), even though the robot never physically touches the screen. In the case of touching objects, we already know the unique identifier of the target object, which allows us to highlight this particular object on the tablet. However, if the task is to move an object to a new location, the desired target location was previously not known to the system. Therefore, the storyboards were extended such that, in the definition of a move task, the coordinates of a correct target location would have to be included:

```
<giveResponseToMoveObject(
    bucket,
    NOT_above,
    shelf,
    -133.40874719142187,
    0,
    73.25956648843038
)>
```

In this task, the child is asked to remove a bucket from the shelf. The last three values are the X, Y and Z coordinates of a location on the screen that satisfies this task. This is the location that the robot would move it to when providing help, although there are many other locations that also satisfy the task condition (and the child is still free to use those). To increase the user-friendliness of creating and editing storyboards, future work should include ways to infer correct locations automatically, such

that the description of the task is enough to allow the robot to perform the task without needing fixed coordinates. For example, the system could automatically scan the environment for a free spot (where there are no other objects yet) that satisfies the condition *NOT_above shelf*, removing the need to manually annotate the desired location.

When the task is to repeat a word or sentence, the robot provides help by offering to repeat the word together with the child:

> Let's say it together. 3, 2, 1.. *monkey!*

This was something we had encountered with a shy child during one of the pilot studies, when there was not yet a mechanism for the robot to provide help in place. In this case, the experimenter would offer to repeat the word with the child, and this appeared to help. In the current implementation, even if the child still does not repeat the word, we move on to the next task after this attempt to provide help.

Because at times it is also desirable for the robot to perform an action, not within the context of helping a child perform a task but because it has tasks of its own, we also included in the storyboard the option to include these interactions with the tablet in mid-sentence:

```
That was fun! Let's bake another bread. Now I want to try it!
<move(bag, 4.5835913413192895,94,42.65093742920607,1,false)> There,
a bag of flour, and into the bowl.
```

In this case, right after the robot finishes the sentence *Now I want to try it!*, the bag of flour is moved towards the bowl's coordinates. This animation is allowed to take one second (the fourth parameter of the *move* command) and is not supposed to loop (fifth parameter). Because the animation is always completed within a certain amount of time, it matches the robot's gesture to create the illusion that the robot is actually manipulating the tablet game – objects that are further from their target location simply move faster than those that are already closer to their destination.

3.5 Interruptions

There is a large amount of output from the robot, which is needed to explain the rules of the game and what action is expected of the child, but also to ensure plenty of exposure to L2 words. Objects on the screen are also locked (cannot be moved) while the robot is still explaining a task, to avoid the system losing track of objects' positions and the overall state of the game. However, during the first pilot sessions it became clear that children want to move faster through the interaction, for example by already attempting to complete a task while the robot is still introducing it. Particularly when a task had to be repeated (e.g., "move three trees into the cage, one by one") children knew what had to be done after the first time. We noticed that children became impatient and sometimes even frustrated when the game did not allow them to perform tasks because the robot was still talking. It also made the robot seem less intelligent as it was less aware of its surroundings, particularly of the fact that the child was ready to proceed.

However, sometimes a particular pace has to be imposed because we want all children to have the same minimum amount of exposures to each target word in the L2, and in some cases the children are clearly still guessing what the task will be. To include this knowledge of where the important part of an utterance ends and the robot may be interrupted, together with WP5 we came up with the tag $< accept\_answer >$ that can be placed in the storyboards, anywhere within an utterance before an action is expected from the child. Whenever this tag is encountered during run-time, while the robot is speaking, objects on the tablet are unlocked and the child can start interacting with them. When the

child actually touches the tablet, the robot's speech is interrupted and the InteractionManager starts evaluating whether the performed action is the correct one.

If children do decide to interrupt the robot and perform a task, but incorrectly, negative feedback includes context, as explained in 3.3.2, such that the child should still be able to understand the task even if the original introduction was interrupted. As it is implemented now, the tag needs to be present somewhere in the text whenever the child is required to interact with the tablet. A straight-forward extension of the system could simply start accepting answers at the end of the robot's utterance, if this tag was not found within the text. This is already implemented for tasks where the child has to repeat a word or sentence, where the robot cannot be interrupted because we anticipate that a future speech recognition implementation will not be able to distinguish between the voice of the child and the robot.

3.6 Logging

With our colleagues that are responsible for the other modules, we agreed on a strategy of logging everything that happens during the interaction. Each child has a memory file, where all important events are stored per lesson. This file also contains information that is used to start the interaction, such as the child's name, identifier and the experimental condition. During each lesson, new events are appended such that, at the end of all lessons, the file provides a complete overview of the child's interactions with the system. The InteractionManager is in charge of these files, while other modules can send their events to be stored. For example, task performance (answer given, whether it was correct) is provided by the InteractionManager. The OutputManager provides several events that are linked to the main functionalities that the module has to offer, which will help to identify patterns that occurred during the interaction, as well as the number of exposures to the L2 target words and to iconic gestures:

- Feedback given: the number of times feedback was previously given, whether it is positive or negative, which template was used and if there are any exposures to target words (currently only with negative feedback)
- Gesture produced: whether it was an iconic gesture (could also be pointing or moving) and the identifier of the gesture
- Request for answer: the number of times an answer was requested for this task already, and if there are any exposures to target words
- Providing help: if there are any exposures to target words (currently only happens when the task is to repeat something)

These events are all linked to points in time within the interaction, so that after aligning the starting points of video and memory the data should be synchronized for the duration of the lesson. Each module also has its own file where more detailed logs are placed, in the case of the OutputManager this includes for instance every utterance said by the robot. The goal of these logs is mostly debugging and they act as a fall-back in case something should happen to the memories.

3.7 Tablet speech

One of the limitations of the robot's text-to-speech engine is that the pronunciation is not always clear and correct. This is a challenge, especially when the goal of the interaction is to teach exactly those words that are being mispronounced. To mitigate this, within the project we decided to ensure that the

first exposure to a new target word or sentence would always have to come from a recording of a native speaker. Usually, this recording is triggered after the child is asked to interact with an object on the screen, so that it feels like a natural response from the tablet game:

> ...And I think we have four bags of flour. [...] Touch them, so that we can learn what four is in English.

After this, subsequent exposures will come from the robot so the learner is in fact exposed to two different versions of the target word. The drawback to this approach is in its scalability: when designing additional lesson content that includes new target words, samples of the pronunciation of the words by a native speaker have to be recorded. This has currently been done for all of the target words (and some sentences that include these words) from the large-scale study, where we recorded a native (Canadian) speaker of English. Several examples of each word were recorded so that the best (most clear) variation could be extracted. The best sample of each word was extracted, noise was removed and it was then saved to a file, which is placed in a specific directory that is part of the output module. The file names of recordings can then be inserted into the storyboards at the desired location, so that they will be played back during the lesson.

4 Ongoing and planned research

4.1 Large-scale study

Data collection for the large-scale study as proposed in the revised objectives document is currently ongoing. The research question and main hypotheses have been preregistered at AsPredicted [12]. From the perspective of WP6, we are particularly interested in the differences between the learning outcomes of children in the robot conditions with and without iconic gestures (H3 of the preregistration). Based on existing research, as well as our findings from our previous experiment, we expect children to retain more English words over time when iconic gestures were present while learning these words. Other than trying to reproduce the positive effects of gestures we discovered previously, there are several differences between our previous study and this large-scale study that could lead to new insights regarding the design and application of gestures in a tutoring system. For example, the positioning of the robot relative to the child has changed. During the previous experiment the robot was standing opposite the child, approximately a meter away, and performing full body gestures while the child was sitting. In the large-scale study both child and robot are sitting down, close to each other, and the robot is on the right-hand side of the child. The gestures have therefore become “smaller”, consisting of only arm and hand movements. This appears to affect the interaction in various ways: children are more tempted to start touching the robot, and gestures are re-enacted more often than in our previous set-up. Furthermore, the effect of iconic gestures may also be affected by the fact that there are now multiple sessions instead of one, that the lessons are more complex and less repetitive, and that there is a larger number of words to learn (34 compared to 6).

4.2 Comprehensibility of recorded gestures

In our previous experiment with iconic gestures, as well as the large-scale study, the gestures that were performed by the robot were all designed manually. This was done by positioning the robot’s limbs and storing the orientation of the robot’s joints as key frames. When playing back these gestures, the robot then interpolates between these key frames to create a motion. The main drawback of this method is

that the implementations of the gestures are based upon the ideas and mental models of the person that is designing them, combined with the physical limitations (e.g., limited degrees of freedom) of the robot. We feel that a more natural way to design these gestures is to have the robot learn them from humans, performed in a setting that is as natural and spontaneous as possible. This learning process can be done using a depth sensor such as Kinect to record gestures, followed by a mapping that translates these human joint positions into joint orientations specific to a robotic platform, in our case the Softbank Robotics NAO. Because the Kinect recordings are unaffected by this mapping process, they can easily be used to also map to different platforms as new, different robots become available.

When trying to record and then map gestures onto the robot, details will be lost due to the limitations of the robot as well as the recording hardware. We intend to investigate to what extent this loss of information reduces the comprehensibility of the gesture, by conducting a study where participants are shown recordings of gestures being performed (by a human or a robot) and are asked to guess which object is being depicted.

5 Conclusion and Future Directions

In this deliverable we have described the current status of the output generation module of the L2TOR system. Recent additions to this system include the implementation of iconic gestures, a variation of the system that runs without robot (tablet only), context-sensitive language generation and coordination with the tablet game. The tutoring system is currently being used for the large-scale study described in section 4.1. Deliverable 6.3 will discuss our findings from this study.

As mentioned in D6.1 with respect to the (automatic) mapping of human gestures to humanoid robots, we are currently preparing an experiment to investigate to what extent gestures can be extracted from video recordings of gestures from human participants to be mapped to the NAO robot with limited (or possibly without) supervision. An important motivation for this study is to find how well these gestures can be mapped to the robot and whether their intended meaning is preserved in the process. In addition, we intend to examine the gesture recognition capabilities of a robot system supported by a Kinect camera.

References

[1] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

[2] Emiel Krahmer, Sebastiaan van Erk, and André Verleg. Graph-based Generation of Referring Expressions. *Computational Linguistics*, 29(1):53–72, March 2003.

[3] Emiel Krahmer and Kees van Deemter. Computational Generation of Referring Expressions: A Survey. *Computational Linguistics*, 38(1):173–218, March 2012.

[4] Candace L Sidner, Cory D Kidd, Christopher Lee, and Neal Lesh. Where to look: a study of human-robot engagement. In *Proceedings of the 9th international conference on Intelligent user interfaces*, pages 78–84. ACM, 2004.

[5] Sean Andrist, Tomislav Pejsa, Bilge Mutlu, and Michael Gleicher. Designing effective gaze mechanisms for virtual agents. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 705–714. ACM, 2012.

[6] Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. The effect of a robot's gestures and adaptive tutoring on children's acquisition of second language vocabularies. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 50–58. ACM, 2018.

[7] Marion Tellier. The effect of gestures on second language memorisation by young children. *Gestures in Language Development*, 8(2):219–235, 2008.

[8] Spencer D. Kelly, Tara McDevitt, and Megan Esch. Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, 24(2):313–334, 2009.

[9] Manuela Macedonia, Karsten Müller, and Angela D. Friederici. The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping*, 32(6):982–998, 2011.

[10] Meredith L. Rowe, Rebecca D. Silverman, and Bridget E. Mullan. The role of pictures and gestures as nonverbal aids in preschoolers' word learning in a novel language. *Contemporary Educational Psychology*, 38(2):109 – 117, 2013.

[11] Jacqueline A. de Nooijer, Tamara van Gog, Fred Paas, and Rolf A. Zwaan. Effects of imitating gestures during encoding or during retrieval of novel verbs on children's test performance. *Acta Psychologica*, 144(1):173 – 179, 2013.

[12] Rianne van den Berghe, Mirjam de Haas, Emiel Krahmer, Paul Leseman, Ora Oudgenoeg, Josje Verhagen, Paul Vogt, Bram Willemsen, and Jan de Wit. Aspredicted.org preregistration — L2TOR (#8181). http://web.archive.org/web/*/https://aspredicted.org/6k93k.pdf. Accessed: 03-05-2018.

The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies

Jan de Wit
TiCC*
Tilburg University
j.m.s.dewit@uvt.nl

Thorsten Schodde
Faculty of Technology, CITEC‖
Bielefeld University
tschodde@techfak.uni-bielefeld.de

Bram Willemsen
TiCC*
Tilburg University
b.willemsen@uvt.nl

Kirsten Bergmann
Faculty of Technology, CITEC‖
Bielefeld University
kirsten.bergmann@uni-bielefeld.de

Mirjam de Haas
TiCC*
Tilburg University
mirjam.dehaas@uvt.nl

Stefan Kopp
Faculty of Technology, CITEC‖
Bielefeld University
skopp@techfak.uni-bielefeld.de

Emiel Krahmer
TiCC*
Tilburg University
e.j.krahmer@uvt.nl

Paul Vogt
TiCC*
Tilburg University
p.a.vogt@uvt.nl

ABSTRACT

This paper presents a study in which children, four to six years old, were taught words in a second language by a robot tutor. The goal is to evaluate two ways for a robot to provide scaffolding for students: the use of iconic gestures, combined with adaptively choosing the next learning task based on the child's past performance. The results show a positive effect on long-term memorization of novel words, and an overall higher level of engagement during the learning activities when gestures are used. The adaptive tutoring strategy reduces the extent to which the level of engagement is diminishing during the later part of the interaction.

KEYWORDS

Language tutoring; Robotics; Education; Human-Robot Interaction; Bayesian Knowledge Tracing; Non-verbal communication

ACM Reference Format:
Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. 2018. The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies. In *HRI '18: 2018 ACM/IEEE International Conference on Human-Robot Interaction, March 5–8, 2018, Chicago, IL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3171221.3171277

1 INTRODUCTION

Robots show great potential in the field of education [24]. Embodied agents in the form of humanoid robots, in particular, may deliver educational content for various subjects in ways similar to human tutors. The main advantage of using such a robot compared to traditional learning tools is its physical presence in the referential world of the learner [20]. The human-like appearance and presence in the physical environment may facilitate interactions that are, to some extent, similar to the ways in which human teachers would communicate with their students. Care should be taken, however, to design for the correct amount of social behavior, so as to avoid distracting students from the task at hand [16].

When designing such interactions, we can draw upon ways in which human teachers give contingent support to students in their learning activities. For instance, particularly in one-on-one tutoring situations, teachers tend to adjust the pace and difficulty of learning tasks based on the past development and current skill set of the student [29]. For example, teachers may help by scaffolding, taking the initial knowledge base as a starting point and trying to optimize the learning gain by choosing the hardest task to perform that still lies within the zone of proximal development [32] of the student.

The use of gestures that coincide with speech is another way for teachers to provide scaffolding, particularly when the concepts which the gestures refer to are not yet mastered by the student [1]. For instance, when teaching a second language (L2), gestures can help to ground an unknown word in the target language by linking it iconically or indexically to a real world concept. Such a facilitating effect on word learning has been found for imitating gestures of a virtual avatar [2]. However, it is an open question if the embodied presence of a robot can be exploited to support language learning through a robot's gesturing, and if so, what kind of gestures would have a positive impact.

In this paper, we present the results of an experiment conducted to explore how these two tools for scaffolding the learning of language — choosing the task that yields the greatest potential learning gain for a particular student and the use of appropriate co-speech gestures — carry over to a humanoid robot. Both were combined

HRI '18, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-4953-6/18/03…$15.00
https://doi.org/10.1145/3171221.3171277

*Tilburg center for Cognition and Communication
‖Cluster of Excellence Cognitive Interaction Technology

in one study to better estimate what the relative importance of the respective techniques is, while keeping all other factors constant, and to find out whether the benefits of the two strategies can potentially reinforce or impede each other. The techniques were implemented and tested in a one-on-one tutoring system where children, four to six years old, play a game with a robot to learn an L2. In the next section, we briefly present the approaches taken to realize the adaptive tutoring along with co-speech gesturing of the robot. We then describe the experimental methodology, before reporting and discussing the results obtained.

2 BACKGROUND

2.1 Adaptive Bayesian Knowledge Tracing

A robot tutor that personalizes the learning experience for individual students has been shown to have a positive effect on performance [19]. This robot is also perceived as smarter or more intelligent and less distracting or annoying. In order to simulate the way human tutors tailor learning activities and difficulty levels to a particular student, an adaptive tutoring system would have to measure and track the knowledge level of the student. Often the knowledge is traced skill-wise, where in the case of language learning, the mastery of particular words or phrases in the target language is represented probabilistically (e.g., [11]). This approach yields promising results, but it lacks flexibility because of the need to define domain-specific distance metrics to choose the next skill. Others have used Dynamic Bayesian Networks to represent the learner's knowledge about a skill, conditioned on the past interaction and taking into account skill interdependencies [14]. This approach requires detailed knowledge about the learning domain to model those interdependencies and their parameters. Recently, Spaulding et al. [27] used a simpler approach based on Bayesian Knowledge Tracing (BKT) [6]. The general BKT model consists of latent variables S^t representing the extent to which the system believes a particular skill to be mastered by the student. The belief state of the system is updated based on observed variables O^t, which correspond to the result of a learning action (e.g., correctly or incorrectly answering a question), while accounting for possible cases of guessing $p(guess)$ and slipping $p(slip)$ during the answer process. It was shown that this model outperforms traditional approaches for tracing the knowledge state in learning interactions, and that it can be easily extended to, for example, incorporate the emotional state of a child. In previous work [26], we have extended the basic BKT with action nodes to also model the tutor's decision-making based on current beliefs about the student's knowledge state (see Figure 1). Additionally, we employed a latent variable S that can attain discrete values for each skill, corresponding to six bins for the belief state (0%, 20%, 40%, 60%, 80%, 100%). This allows for quantifying the robot's uncertainty about a learner's skills as well as the impact of tutoring actions on future observations and skills.

This so-called *Adaptive* Bayesian Knowledge Tracing (A-BKT) approach can be used to choose the next skill from which the learner will most likely benefit, by estimating the greatest expected knowledge gains. It tries to maximize the belief of each skill while also balancing over all skills and not teaching a particular skill over and over again, even if the answer to the task was wrong and the skill belief is the lowest. The system does not only allow to choose the best skill to address next, but also the action to be used for scaffolding the learning of this skill. In this context, actions can be, for example, different types of exercises, pedagogical acts, or task difficulties. For the sake of simplicity, three task difficulties have been established (easy, medium, hard) to address a skill and to find the best action for a given skill.

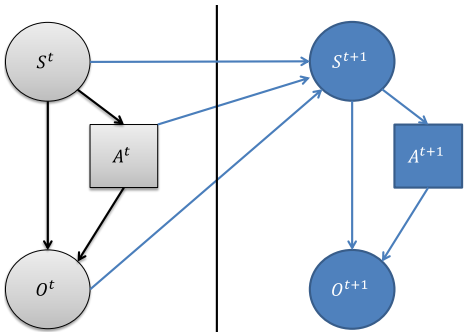


Figure 1: Dynamic Bayesian Network for BKT (taken from [26], with permission): with the current skill-belief the robot chooses the next skill S^t and action A^t for time step t and observes O^t as response from the user.

The goal of this strategy is to create a feeling of flow which can lead to better learning results [7]. It strives not to overburden the learner with tasks that would be too difficult nor to bore them with tasks that would be too easy, both of which may lead to disengagement and thus hamper the learning. Note that this approach is comparable to the vocabulary learning technique of *spaced repetition* as implemented, for instance, in the Leitner system [18]. The implementation of A-BKT used in the current study is identical to the one used previously in [26]. However, it has not yet been evaluated with children nor in conjunction with other techniques that might affect action difficulty (such as gestures). Furthermore, its impact on student engagement has not been explored previously.

2.2 Gestures

Iconic gestures elicit a mental image that corresponds directly, either in form or execution, to the concept or action that is being described verbally at the same time [23]. For example, a flying bird could be depicted by stretching both arms sideways and moving them up and down. Studies have shown that iconic gestures, when performed by a human teacher, may aid the acquisition of L2 vocabularies [8, 15, 21, 28]. Hald et al. [12] provide an overview of how gestures can contribute to learning an L2. They propose that gestures might have a 'grounding' effect by linking existing perceptual and motor experiences to a new word. This is expected to result in a richer mental representation. Research by Rowe et al. [25] shows that gender, language background, and level of experience in the native language (L1) influence the extent to which gestures can contribute to L2 learning. The positive effects of gestures hold true for younger students as well; in fact, gestures are suggested to be a crucial part of communication with children [13]. It has also been

shown that gestures help not only to acquire knowledge, but also to retain it over time [5].

Previous research has explored the use of gestures by virtual agents (e.g., [2]) and robots (e.g., [30]), finding similar, positive effects on memory performance when gestures are produced by an artificial embodied agent compared to a human tutor. While humans tend to spontaneously perform and time their gestures, they will often need to be manually designed and coordinated with speech for the robot. Due to its limited degrees of freedom, however, the robot is unable to perform motions with the same level of detail, finesse, and accuracy as a human. This may lead to a loss in meaning when human gestures are being translated directly to the robot, indicating a need for alternative gestures. As a concrete example, the SoftBank Robotics NAO robot that was used in this case is unable to move its three fingers individually, preventing it from performing pointing gestures or finger-counting. However, research suggests that iconic gestures are almost as comprehensible when performed by a robot, compared to a human [4].

3 METHODOLOGY

An experiment was conducted to investigate the effect of using iconic gestures and an adaptive tutoring strategy on children's acquisition of L2 vocabularies, with the intention of answering the following three hypotheses:

> H1: There is a greater learning gain when target words are accompanied by iconic gestures during training, than in the case of not using gestures.
> H2: There is a reduced knowledge decay when target words are accompanied by iconic gestures during training, than in the case of not using gestures.
> H3: There is a greater learning gain when target words are presented in an adaptive order during training, based on the knowledge state of the child, than when target words are randomly introduced.

These hypotheses rely upon the underlying assumption that children are able to acquire new L2 words during a single session with a robot tutor, regardless of experimental conditions; this assumption was also put to the test.

The experiment had a 2 (adaptive versus non-adaptive) x 2 (gestures versus no gestures) between-subjects design. In the two conditions with the adaptive tutoring strategy, the A-BKT system described in Section 2.1 was used to select the target word for each round, based on the believed knowledge state of the child. In practice, this meant that children would be presented with a particular target word more frequently if they had answered it incorrectly in the past, thereby changing the number of times each target word occurred during training, although each target word was guaranteed to occur at least once. Other conditions had a random selection, where each of the six target words would always be presented five times, in a randomized order, for a total of thirty rounds. In the gesture conditions, whenever a target word was introduced in the L2 it was accompanied by an iconic gesture (as shown in Figure 2). All conditions had the robot standing up and in "breathing" mode, which meant that it slowly shifted its weight from one leg to the other and had a slight movement in its arms to simulate breathing.

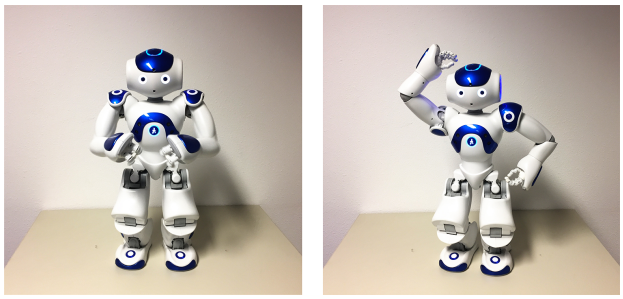


Figure 2: Examples of the stroke of two iconic gestures performed by the robot (taken from [9], with permission). Left: imitating a *chicken* by simulating the flapping of its wings; right: imitating a *monkey* by scratching head and armpit.

3.1 Participants

Participants were 61 children, with an average age of 5 years and 2 months ($SD = 7\,months$), 32 girls. They were recruited from primary schools in the Netherlands, by first contacting schools and then sending out an information letter together with a consent form through the schools to the parents of children that satisfied the age limit of four to six years. Only native Dutch children with Dutch as their L1 are included in the evaluation, although all 99 children that had signed up were allowed to participate in the experiment. The children were randomly assigned to conditions, while taking into account a balance in age and gender.

3.2 Materials

The aim of the tutoring interaction was to teach children six animal names in English: bird, chicken, hippo, horse, ladybug, and monkey. These specific words were chosen because the Dutch words are distinctly different from their English translations and because it was possible to create uniquely defining iconic gestures for them.

The SoftBank Robotics NAO robot was used, which was standing in front and slightly to the right of the child. After an experimenter had filled in the name of the child and pressed the start button, the experiment ran fully autonomously. Two experimenters were always present, where one would take care of getting the child from the classroom and explaining the procedure of the experiment, while the other would set up the system. To avoid having the child seek them out for feedback, the experimenters would announce that they would be occupied. The child was asked to sit on pillows, close to the tablet which was raised on a box and slightly tilted. Two cameras were used to record the interaction, one facing the front of the child and one at an angle from the side. The basic setup is shown in Figure 3, although it differed slightly between locations due to the layout of the rooms. In the condition with gestures every occurrence of the target word in L2, except when giving feedback, was accompanied by the matching iconic gesture (see Figure 2). The gesture was timed in such a way that the pronunciation of the target word would coincide with the stroke of the gesture, i.e., the accented phase that is most related to the meaning. A perception study was conducted to evaluate the quality of the gestures [9], where 14 participants were shown video recordings of all six gestures

Figure 3: The setup for the experiments.

performed by the robot and then asked to indicate which out of the six target words corresponds to each particular recording. Based on the results of this study, each gesture was deemed to be sufficiently unique to distinguish between the six target words.

The adaptive tutoring system starts with medium (0.5) confidence for all target words, a value associated with two distractors during training. Each distractor is a false answer to a task, an image belonging to one of the five other target words. In the random conditions, since there is no knowledge tracing the difficulty was always set to medium (two distractors). The tablet was used to get input from the child, because speech recognition does not work reliably with children [17]. This is also why only comprehension and not production of the target words is evaluated. An example of what the tablet screen would look like is shown in Figure 5. The images used during training belong to a different set of images than the ones used for the pre-test and post-tests. The set of images used during training matches the gesture that the robot performs related to the animals, for example the image of the horse for the training stage (shown in Figure 5) also includes a rider because the robot shows the act of riding a horse as a gesture. The image that was used during the tests did not include a rider and the horse is standing still, facing the opposite direction (shown in Figure 4). In addition to changing the pose or context of the animals, colors also varied. Together with having a recorded voice in the tests instead of the robot's synthesized speech, this aims to verify whether children learn how the English words map to the concepts of the animals and their matching Dutch words, rather than to one specific image.

3.3 Procedure

Prior to partaking in the experiment, participants were introduced to the robot during a group introduction. This approach is inspired by the work of Vogt et al. [31] with the intention of lowering the anxiety of children in subsequent one-on-one interactions with the robot. The introduction consisted of a description of what the robot is like, including a background story and how it is similar to humans in some respects, and different in others. Together with the children (and sometimes teachers and experimenters) the robot performed dances, after which all children were presented with

Figure 4: The pre-test and post-tests on a laptop, using a recorded voice and a different set of images from those on the tablet.

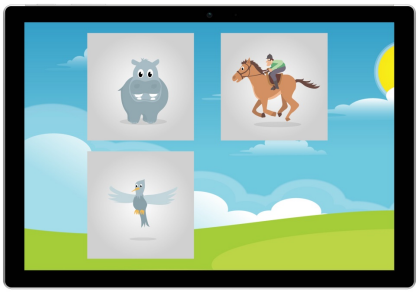


Figure 5: The tablet during training, showing images corresponding to the target word and two distractors.

the opportunity to shake the robot's hand before putting it to bed. Introductory sessions were scheduled several days before the first participant was to take part in the experiment, allowing time for the children to process these new impressions.

Before starting the tutoring interaction, a pre-test was administered to gauge the level of prior knowledge with respect to the animal names in the L1 (Dutch) and L2 (English). This test was administered on a laptop, where images of all six animals were randomly positioned on the screen. A recording of a (bilingual) native speaker pronouncing one of the six animal names was played, after which the child was asked to click the corresponding image on the screen (Figure 4). This was done for all six target words, first in Dutch and then in English.

After completing the pre-tests, the child would go through each target word one by one, still using the laptop. This is done to give the children a first exposure to the correct mappings between target words and the concepts they refer to, to avoid turning the first rounds of learning with the robot into a guessing game. Because there is no feedback during the pre-tests, this also ensures that concepts are linked to the correct word, rather than having the child assume that their answers during the pre-tests were all correct. For each word, the image of the corresponding animal would be shown in the center of the screen and the laptop would play a recording by a (bilingual) native speaker saying: "Look, this is a [target in L2]. Do you see the [target in L2]? Click on the [target in L2]!"

The training stage of the experiment consisted of the child and robot playing thirty rounds of the game *I spy with my little eye*. The robot, acting as the spy, would pick one of six target words and call out: "I spy with my little eye...", followed by the chosen word in the L2. For this stage, children were assigned to one of four conditions:

(1) Random tutoring strategy, no gestures ($N = 16$)
(2) Random tutoring strategy, gestures ($N = 14$)
(3) Adaptive tutoring strategy, no gestures ($N = 15$)
(4) Adaptive tutoring strategy, gestures ($N = 16$)

Prior to playing the game, the robot explained the procedure and asked the child to indicate whether they understood by pressing either a green or a red smiley. If the red smiley is pressed, the interaction would pause and an experimenter would step in to provide any further explanations. After this introduction, there were two practice rounds: one in Dutch and one in English.

After the robot had "spied" an animal, a corresponding image was shown on the tablet along with a number of distractor images (Figure 5). The child was then asked to pick the image that matched the animal name that the robot had spied. The number of distractors was determined by the difficulty level of the round, which in the case of the adaptive conditions depended on the confidence that the system had in that the child knew this particular target word. A low confidence resulted in only one distractor, while a high confidence had three distractors.

Feedback to the task was given by both the tablet and the robot. The tablet highlighted the image selected by the participant, either with a green, happy smiley if the correct answer was provided or a red, sad smiley if the selected image was an incorrect answer. The robot then provided verbal feedback, which in the case of a correct answer consisted of a random pick out of six positive feedback phrases (e.g., "well done!"), followed by "The English word for [target in L1] is [target in L2]". In the case of negative feedback, the robot would say "That was a [chosen answer in L1], but I saw a [target in L2]. [Target in L2] is the English word for [target in L1]". Whenever an incorrect answer was given, the same round would be presented once more but at the easiest difficulty (with only one distractor: the image that was incorrectly chosen in the previous attempt). This, combined with additional exposures in the corrective feedback, means that the number of times each target word was presented in the L2 may vary between children, depending on how many rounds were answered incorrectly. After finishing thirty rounds of training with the robot, the child was asked to complete a post-test on the laptop. This test is identical to the pre-test that was administered at the start of the experiment, in L2. Finally, the post-test was repeated once more, at least one week after the experiment, to measure long-term retention of the newly acquired knowledge.

3.4 Analysis

Immediate learning gain was measured as the difference between the number of correct answers on the post-test, administered directly after the training stage, and the number of correct answers on the pre-test, taken prior to the tutoring interaction. Test scores were always between 0 and 6 because each target word was asked once in the L2. The post-test was administered once more, (at least) one week after the experiment. We then looked at the difference between this delayed test and the pre-test for long-term learning gain. Finally, we took the difference between the delayed test and the immediate post-test as a measure of knowledge decay. The design of these tests is described in more detail in Section 3.2.

Children's tasks during training were of varying task difficulty in the adaptive tutoring condition, with one to three distractor images. To account for these differences, as well as to allow a comparison with the post-test results (five distractor images), we mapped binary task success (1: correct response; 0: incorrect response) onto the span between 0.0 and 1.0 by subtracting a value of 0.2 for each of the potential five distractor images that was not provided, which would, for example, result in a score of 0.6 for a correct response in a task with three distractors. The total score during training was then divided by the number of rounds (30), resulting in a training performance value between 0.0 and 1.0 (Figure 7).

4 RESULTS

The average duration of the training stage of the experiment was 18:38 minutes (SD = 3:03). Including the introduction, pre-test, and post-test this amounted to a session length of roughly thirty minutes. To confirm whether children managed to learn any new words from a single tutoring interaction, regardless of strategy or the use of gestures, a paired-samples t-test was conducted to measure the difference between post-test and pre-test scores for all conditions combined. There was a significant difference between the scores on the pre-test ($M = 1.75$, $SD = 1.14$) and immediate post-test ($M = 2.85, SD = 1.61$), $t(60) = 5.23, p < .001$. The same analysis was conducted for the delayed post-test that was taken (at least) one week after the experiment. Results revealed a significant difference between the pre-test scores ($M = 1.75, SD = 1.14$) and the delayed post-test test scores ($M = 3.02, SD = 1.40$), $t(60) = 6.81, p < .001$. However, there was no significant difference between the delayed post-test and the immediate post-test, $t(60) = .92, p = .34$. This means that H2 is not supported by these results, since no significant decay was observed in any of the conditions.

To investigate the effects of the different conditions on training performance, a two-way ANOVA was carried out with tutoring strategy (adaptive versus non-adaptive) and the use of gestures (gestures versus no gestures) as independent variables and performance during training as the dependent variable (Figure 7). As described in Section 3.4, these scores are weighted by the number of distractors present and divided by 30 rounds, resulting in a value between 0.0 and 1.0. For the 30 rounds of training there was a main effect of gesture use, $F(1, 57) = 18.23, p < .001, \eta_p^2 = .24$, such that training with gestures led to higher score ($M = .38, SD = .09$) than learning without gestures ($M = .29, SD = .08$). Children in the adaptive condition achieved a higher score ($M = .36, SD = .12$) than children in the non-adaptive condition ($M = .32, SD = .06$), but the effect of tutoring strategy was not significant, $F(1, 57) = 3.62, p = .06, \eta_p^2 = .06$. There was a significant interaction effect between use of gestures and tutoring strategy, $F(1, 57) = 4.72, p = .03, \eta_p^2 = .08$. Without gesture use, there was no significant difference between tutoring strategies. When gestures were present, however, children in the adaptive condition turned out to perform better than those in the non-adaptive condition. Hence, children's learning outcome was best when gesture use and adaptive training were combined.

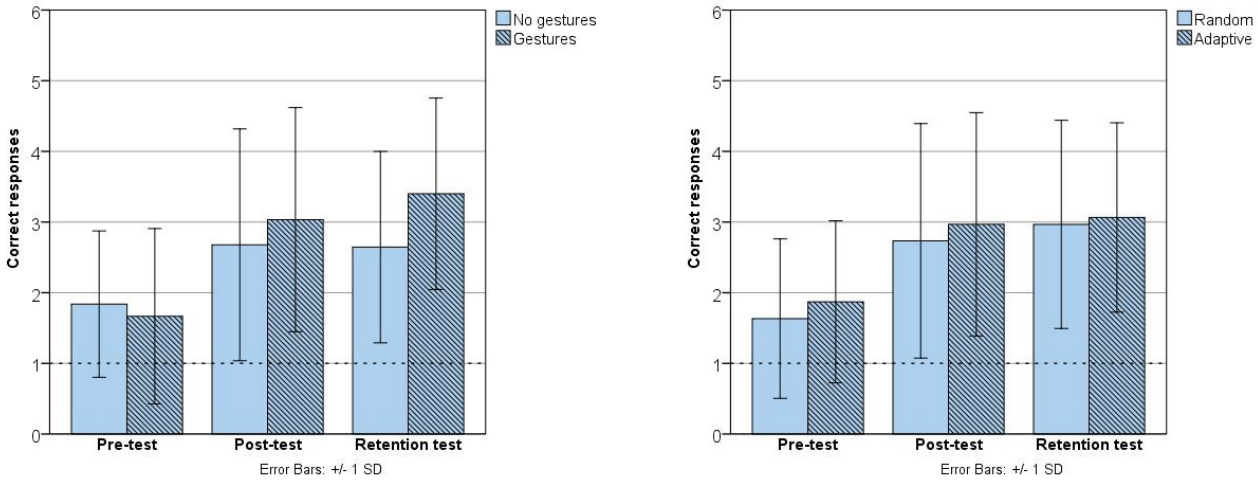


Figure 6: Test scores for the gesture vs no gesture conditions (left) and the adaptive vs random conditions (right).

Another two-way ANOVA was carried out to measure learning gain, with the difference score between the post-test results and the pre-test results as the dependent variable (Figure 6). There was no significant effect of tutoring strategy, $F(1, 57) < .001, p = .95, \eta_p^2 < .001$, or use of gestures, $F(1, 57) = 1.53, p = .22, \eta_p^2 = .03$. These results do not support H1 and H3 (greater learning gains when gestures and adaptive tutoring are used). The same two-way ANOVA with the difference score between results of the delayed post-test and the pre-test also did not give a significant effect of tutoring strategy, $F(1, 57) = .36, p = .55, \eta_p^2 = .006$, but there was a significant effect for use of gestures, $F(1, 57) = 6.11, p = .02, \eta_p^2 = .097$, indicating that the learning gain between pre-test and delayed post-test was greater when gestures were used during training ($M = 1.70, SD = 1.56$) than when no gestures were used ($M = .81, SD = 1.25$). Although this does not fully support H1 or H2, it does show a long-term learning gain when gestures are used during learning. No interaction effect was found, $F(1, 57) = .04, p = .84, \eta_p^2 \leq .001$.

4.1 Evaluation of engagement

The engagement of the children during the training stage with the robot was examined to find out whether children became more disengaged with the tutoring tasks towards the end of the thirty rounds, and whether the application of an adaptive tutoring strategy and gestures would influence the change in engagement levels. This was done by asking 18 adult participants, without specific training in working with children, to rate video clips (without audio) of the children interacting with the robot. The choice for conducting a perception study with adults using video recordings of the experiment was made for two reasons: so that the training would not have to be interrupted for questions regarding the experience, thereby potentially influencing the engagement, and because it is difficult for children of a young age to reflect upon their experiences and verbalize these thoughts [22]. For each child, one clip was taken from the fifth round of training and one clip from the twenty-fifth round, to get observations that are close to the beginning and end

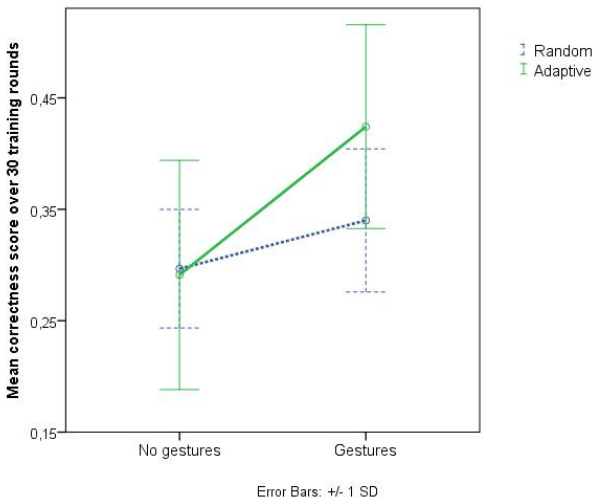


Figure 7: Interaction effects of gesture use and training strategy.

of the training, but far enough from these actual moments to avoid short bursts of engagement when children realize the experiment is starting or finishing. The clips start right after the robot finishes introducing the task, i.e., the point at which the turn switches to the child to provide an answer. All clips then run for five seconds. One child that was excluded from the previous analysis because delayed post-test results were missing, was included for this part of the evaluation. However, data from one other child was missing, making the number of stimuli 122 (61 children, two clips each), with 14 to 16 children in each condition. Participants in the evaluation were asked to rate all 122 clips, randomly presented to them, on a scale from 1 (completely disengaged) to 7 (completely engaged). As a practice round, two clips of a child that was not included in the

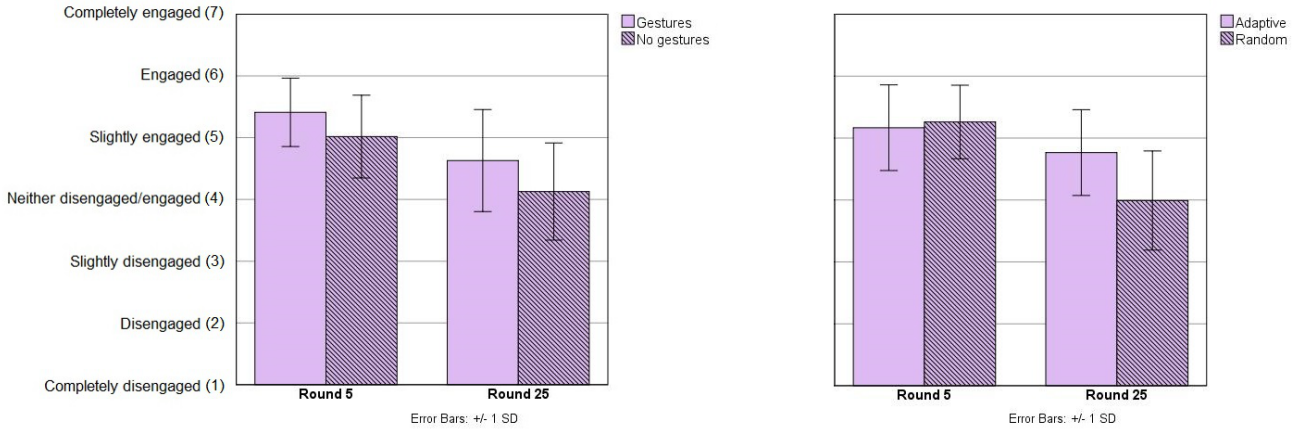


Figure 8: Rated engagement levels early and late in the training interaction for the gesture versus no gesture conditions (left) and the adaptive versus random conditions (right).

main experiment were presented, where one example was clearly engaged and the other was clearly not engaged. After this practice round, participants were told which features from the examples showed engagement (i.e., rapid response to the question, upright body posture, displaying joy after answering the question) and disengagement (i.e., slower response to the question, supporting the head by leaning on the arms, showing less interest in the task).

For each participant, the ratings were averaged over all children belonging to the same experimental condition, resulting in a total of eight average ratings (four conditions, each with fifth and twenty-fifth round). Figure 8 visualizes the data from the evaluation. Results from a paired-samples t-test showed that children were considered to be significantly less engaged in the twenty-fifth round ($M = 4.38, SD = .84$) than in the fifth round ($M = 5.21, SD = .64$), $t(71) = -12.09, p < .001$. Furthermore, a two-way ANOVA with tutoring strategy (adaptive versus non-adaptive) and gesture use (gestures versus no gestures) as factors showed no significant effect for the use of gestures, $F(1, 68) = 1.36, p = .25, \eta_p^2 = .02$, but there was a significant effect for tutoring strategy, $F(1, 68) = 86.26, p < .001, \eta_p^2 = .559$. The drop in engagement between round five and round twenty-five was less when an adaptive strategy was applied ($M = -.40, SD = .35$) than when words were randomly presented ($M = -1.27, SD = .44$). There was no interaction effect between gestures and tutoring strategies, $F(1, 68) = .01, p = .93, \eta_p^2 = .00$. The same analysis was conducted with the average engagement level of the fifth and twenty-fifth rounds combined, to get an idea of the overall engagement throughout the entire training session in different conditions. In this case the overall level of engagement was significantly higher in the gesture condition ($M = 5.02, SD = .63$) than in the condition without gestures ($M = 4.57, SD = .68$), $F(1, 68) = 8.75, p = .004, \eta_p^2 = .114$. There was also a significantly higher engagement when an adaptive strategy was used ($M = 4.97, SD = .67$) as opposed to a random tutoring strategy ($M = 4.63, SD = .67$), $F(1, 68) = 5.10, p = .03, \eta_p^2 = .07$. No interaction effect between the two factors was found, $F(1, 68) = .08, p = .78, \eta_p^2 = .001$.

5 DISCUSSION

The results presented above show that by spending a single tutoring interaction of about twenty minutes with a robot tutor, young children were able to acquire new words in an L2, regardless of the experimental condition, and were also able to retain this newly acquired knowledge for a prolonged period of time. Care was taken to design the pre-test and post-tests in such a way to be clearly distinct from the training session with the robot in terms of physical context (laptop versus tablet), voice, and characteristics of the images used, with the aim of getting a reliable measure of the attained knowledge. Results from the pre-test show that there is indeed a realistic amount of prior knowledge, on average above chance, presumably because some children have been exposed previously to the target words, for example in television programs. The observed number of correct answers on the immediate and delayed post-test are higher than on the pre-test, indicating the expected knowledge gain after engaging in learning activities. The scores on the post-test are lower than the number of correct answers towards the end of the training stage, which could show that indeed the test evaluates whether children acquire the underlying concepts, rather than simply being able to link a word being pronounced by the robot to one specific image (in some cases with the help of gestures that are not present in the tests). One potential point of improvement for the tests could be to introduce context when querying the target words, for example by using sentences rather than isolated words. Although explicitly instructed, children seemed not always aware that they were supposed to select the image corresponding to an *English* word, causing them to choose the animal with the most similar sounding name in Dutch instead (e.g., bird was often confused with the Dutch word 'paard').

When gestures were performed by the robot during training, there was a higher retention of newly acquired words after at least one week. This aligns with similar effects that were shown previously in the context of math with a human tutor [5] and indicates that these indeed carry over to a robot; a compelling finding that warrants future research into the intricacies of gesture use by humanoid robots. As mentioned by Hostetter [13] with respect to

human-human communication, it appears that gestures retain their positive effects on communication when they are scripted rather than being produced spontaneously. In this work, only iconic gestures are used that clearly relate to the concept they describe. Future work could investigate whether a similar contribution to learning gain is found when non-iconic gestures are used. Furthermore, the target words used in this experiment were chosen specifically such that matching gestures could be designed for the robot. It would be interesting to explore how well a broader range of gestures, describing various abstract and concrete concepts, could be performed by a robot as opposed to a human interlocutor. Finally, asking children to actually re-enact the gestures (e.g., as in [8, 28]), or to come up with their own gestures, might further increase the potential utility of gestures in learning due to the embodiment effect [10].

The test results regarding the adaptive tutoring system are currently inconclusive. This might be a result of the manner in which learning gain was measured, i.e., a quantification of newly acquired words — perhaps the adaptive system did not result in *more* words learned, but rather led to a more focused acquisition of exactly those words that the child found most difficult. The main remaining difference between the ways in which human teachers and the system presented here personalize content is that teachers tend to draw upon a memory that spans a longer period of time. In this experiment, the memory of the adaptive system was built up, and then applied, over the course of a single session. The system might come to fruition if there are multiple sessions with the same child, allowing the results of one session to become prior knowledge for the next one. It is also possible that the actions that the system performs based on the estimated knowledge levels of the child are too subtle. Currently, only the order and frequency of words is tailored, within the thirty rounds, and different levels of difficulty are represented by adding or removing one distractor image. Actions and difficulty levels could be more complex than that, for example by applying completely different tutoring strategies or games that might fit a particular child better. For the sake of this experiment, the number of rounds was fixed to thirty, but this session length might also be left up to the adaptive system to control. This would allow the interaction to end at the exact moment where the learning is 'optimal', i.e., a point at which the adaptive system thinks that the child has achieved his or her highest potential learning gain. A final avenue for improvement that is currently being pursued is to incorporate additional information about the affective state of the child. Some children might not be in the right mood to learn when they start, or their attention might fade during the interaction; rather than focusing only on the learning objectives the robot might want to engage in activities that work towards creating and maintaining the right atmosphere for learning.

We found it valuable to include the measure of children's engagement during the interaction. A higher level of engagement indicates increased motivation and willingness to learn [3]. Although students might succeed in simple word learning with limited engagement and the use of a low-level learning strategy, increased engagement could stimulate them to go beyond simple memorization and relate these new words to prior knowledge. Furthermore, engagement can serve as a measure of how well the learning activities are tailored to the child's abilities — constantly presenting tasks that are either too hard or too easy could have a detrimental effect on engagement. The results of our evaluation show that indeed the adaptive system appears to match the learning activities to each child's needs by providing a realistic yet challenging task, resulting in a reduced decline in engagement towards the end of the interaction. Gestures contribute to a higher overall engagement, which could be explained by the fact that the robot appears more active and playful in this condition, thereby stimulating the child to remain engaged.

6 CONCLUSION

The study presented in this paper aimed to explore if a humanoid robot can support children, four to six years old, in learning the vocabulary of a second language. We found that, indeed, children manage to learn new words during a single tutoring interaction, and are able to retain this knowledge over time. Specifically, we investigated whether the effects of tailoring learning tasks to the knowledge state of the learner and using co-speech gestures — both of which are strategies used by human teachers to scaffold learning — transfer to the use of a humanoid robot tutor. Our results show that the robot's use of gestures has a positive effect on long-term memorization of words in the L2, measured after one week. Furthermore, children appear more engaged throughout the tutoring session and are able to provide more correct answers when gestures are used. An adaptive tutoring strategy helps to reduce the drop in engagement that inevitably happens over the course of an interaction, by providing contingent, personalized support to each learner. By combining both methods in a tutoring session, adaptivity seems to succeed in finding the 'sweet spot' of challenging children enough to keep them motivated while gestures can add to overall engagement and support children in finding the correct answer. Therefore, gestures can form an additional tool in the toolbox of A-BKT to be deliberately employed, for example, when a reduced difficulty is deemed necessary or engagement is decreasing.

ACKNOWLEDGMENTS

This work is partially funded by the H2020 L2TOR project (grant 688014), the Tilburg center for Cognition and Communication 'TiCC' at Tilburg University (Netherlands) and the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277), funded by the German Research Foundation (DFG), at Bielefeld University (Germany). The authors would like to thank all members of the L2TOR project for their valuable comments and suggestions that have contributed towards the design of the experiment. Furthermore, we are grateful to the schools, parents, and children that participated in our experiment, Elske van der Vaart for lending us her voice for the content on the laptop, as well as Sanne van Gulik, Marijn Peters Rit, and Emmy Rintjema for their help with data collection. The preliminary design of this experiment was first presented at the R4L workshop, HRI'17 [9]; we thank the attendees for their feedback.

REFERENCES

[1] Martha W. Alibali and Mitchell J. Nathan. 2007. Teachers' Gestures as a Means of Scaffolding Students' Understanding: Evidence From an Early Algebra Lesson. *Video Research in the Learning Sciences* 39, 5 (2007), 349–366. https://doi.org/10.1111/j.1467-8535.2008.00890_7.x

[2] Kirsten Bergmann and Manuela Macedonia. 2013. A virtual agent as vocabulary trainer: iconic gestures help to improve learnersâĂŹ memory performance. In *International Workshop on Intelligent Virtual Agents*. Springer, 139–148.

[3] Phyllis C. Blumenfeld, Toni M. Kempler, and Joseph S. Krajcik. 2005. *Motivation and Cognitive Engagement in Learning Environments*. Cambridge University Press, Cambridge, Chapter 28, 475–488. https://doi.org/10.1017/CBO9780511816833.029

[4] Paul Bremner and Ute Leonards. 2016. Iconic gestures for robot avatars, recognition and integration with speech. *Frontiers in Psychology* 7 (feb 2016), 183. https://doi.org/10.3389/fpsyg.2016.00183

[5] Susan Wagner Cook, Zachary Mitchell, and Susan Goldin-Meadow. 2008. Gesturing makes learning last. *Cognition* 106, 2 (2008), 1047–1058. https://doi.org/10.1016/j.cognition.2007.04.010 arXiv:NIHMS150003

[6] Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[7] Scotty Craig, Arthur Graesser, Jeremiah Sullins, and Barry Gholson. 2004. Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of educational media* 29, 3 (2004), 241–250.

[8] Jacqueline A. de Nooijer, Tamara van Gog, Fred Paas, and Rolf A. Zwaan. 2013. Effects of imitating gestures during encoding or during retrieval of novel verbs on children's test performance. *Acta Psychologica* 144, 1 (2013), 173–179. https://doi.org/10.1016/j.actpsy.2013.05.013

[9] Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. 2017. Exploring the Effect of Gestures and Adaptive Tutoring on Children's Comprehension of L2 Vocabularies. In *Proceedings of the Workshop R4L at ACM/IEEE HRI 2017*.

[10] Katinka Dijkstra and Lysanne Post. 2015. Mechanisms of embodiment. 6, OCT (2015), 1525. https://doi.org/10.3389/fpsyg.2015.01525

[11] Goren Gordon and Cynthia Breazeal. 2015. Bayesian Active Learning-based Robot Tutor for Children's Word-reading Skills. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*. AAAI Press, 1343–1349.

[12] Lea A. Hald, Jacqueline de Nooijer, Tamara van Gog, and Harold Bekkering. 2016. Optimizing Word Learning via Links to Perceptual and Motoric Experience. *Educational Psychology Review* 28, 3 (2016), 495–522. https://doi.org/10.1007/s10648-015-9334-2

[13] Autumn B. Hostetter. 2011. When do gestures communicate? A meta-analysis. *Psychological Bulletin* 137, 2 (2011), 297–315. https://doi.org/10.1037/a0022128

[14] Tanja Käser, Severin Klingler, Alexander Gerhard Schwing, and Markus Gross. 2014. Beyond knowledge tracing: Modeling skill topologies with bayesian networks. In *International Conference on Intelligent Tutoring Systems*. Springer, 188–198.

[15] Spencer D. Kelly, Tara McDevitt, and Megan Esch. 2009. Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes* 24, 2 (2009), 313–334. https://doi.org/10.1080/01690960802365567 arXiv:http://dx.doi.org/10.1080/01690960802365567

[16] James Kennedy, Paul Baxter, and Tony Belpaeme. 2015. The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 67–74. https://doi.org/10.1145/2696454.2696457

[17] James Kennedy, Severin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child Speech Recognition in Human-Robot Interaction : Evaluations and Recommendations. *Proc. of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2017), 82–90. https://doi.org/10.1145/2909824.3020229

[18] S. Leitner. 1972. *So lernt man Lernen: Der Weg zum Erfolg [Learning to learn: The road to success]*. Freiburg: Herder.

[19] Daniel Leyzberg, Samuel Spaulding, and Brian Scassellati. 2014. Personalizing robot tutors to individuals' learning differences. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 423–430.

[20] Daniel Leyzberg, Samuel Spaulding, Mariya Toneva, and Brian Scassellati. 2012. The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains. *34th Annual Conference of the Cognitive Science Society* 34, 1 (jan 2012), 1882–1887. https://doi.org/ISBN978-0-9768318-8-4

[21] Manuela Macedonia, Karsten Müller, and Angela D. Friederici. 2011. The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping* 32, 6 (2011), 982–998. https://doi.org/10.1002/hbm.21084

[22] Panos Markopoulos, Janet C. Read, Stuart MacFarlane, and Johanna Hoysniemi. 2008. *Evaluating Children's Interactive Products: Principles and Practices for Interaction Designers*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, Chapter 1, 3–18.

[23] David McNeill. 1985. So you think gestures are nonverbal? *Psychological Review* 92, 3 (1985), 350–371. https://doi.org/10.1037/0033-295x.92.3.350

[24] Omar Mubin, Catherine J. Stevens, Suleman Shahid, Abdullah Al Mahmud, and Jian-Jie Dong. 2013. A Review of the Applicability of Robots in Education. *Technology for Education and Learing* 1 (2013), 209—-0015. https://doi.org/10.2316/Journal.209.2013.1.209-0015

[25] Meredith L. Rowe, Rebecca D. Silverman, and Bridget E. Mullan. 2013. The role of pictures and gestures as nonverbal aids in preschoolers' word learning in a novel language. *Contemporary Educational Psychology* 38, 2 (2013), 109–117. https://doi.org/10.1016/j.cedpsych.2012.12.001

[26] Thorsten Schodde, Kirsten Bergmann, and Stefan Kopp. 2017. Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making. In *Proceedings of ACM/IEEE HRI 2017*. ACM Press, 128–136. https://doi.org/10.1145/2909824.3020222

[27] Samuel Spaulding, Goren Gordon, and Cynthia Breazeal. 2016. Affect-Aware Student Models for Robot Tutors. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 864–872.

[28] Marion Tellier. 2008. The effect of gestures on second language memorisation by young children. *Gesture* 8, 2 (2008), 219–235. https://doi.org/10.1075/gest.8.2.06tel

[29] Janneke van de Pol, Monique Volman, and Jos Beishuizen. 2010. Scaffolding in teacher-student interaction: A decade of research. (2010), 271–296 pages. https://doi.org/10.1007/s10648-010-9127-6 arXiv:arXiv:1002.2562v1

[30] Elisabeth T. van Dijk, Elena Torta, and Raymond H. Cuijpers. 2013. Effects of Eye Contact and Iconic Gestures on Message Retention in Human-Robot Interaction. *International Journal of Social Robotics* 5, 4 (2013), 491–501. https://doi.org/10.1007/s12369-013-0214-y

[31] Paul Vogt, Mirjam De Haas, Chiara De Jong, Peta Baxter, and Emiel Krahmer. 2017. Child-Robot Interactions for Second Language Tutoring to Preschool Children. *Frontiers in human neuroscience* 11, March (2017), 1–7. https://doi.org/10.3389/fnhum.2017.00073

[32] Lev Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Harvard University Press, Cambridge, MA.