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

D6.1: Output module for number domain

Due Date: **31/06/2017**
Submission Date: **06/08/2017**

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 number domain. We discuss how we addressed challenges regarding multi-modality as well as NLG for a largely scripted interaction. We explain the architecture of the module, highlight its place within the integrated system, outline its functions, and describe the functionality of the various submodules of the OutputManager in more detail. We also briefly touch upon the transformation of the storyboards (Deliverable 2.1) to a machine-readable format. In addition, we include some preliminary findings of a study on the effects of the use of iconic gestures in support of L2 word learning. Finally, we describe the difficulties with respect to natural-sounding speech synthesis and how the phonetic transcription of utterances may resolve some of the problems we have encountered thus far.

Principal Contributors

The main authors of this deliverable are as follows:

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

Laura Hoffmann, Bielefeld University
Amit K. Pandey, SoftBank Robotics
Rianne van den Berghe, Utrecht University

Revision History

Version 1.0 (05-08-2017)
First version.

1 Introduction

When interacting with a child, the robot's communicative behaviour is realized by the output generation module. Here, planned actions are aggregated and executed in a way that is expected to further the interaction between the robot and its conversational partner. The output generation module, developed as part of WP 6, is responsible for deciding how to best execute the robot's next action, based on the input of the different modules of the system. The robot is capable of addressing its conversational partner in spoken natural language. Utterances will be generated using templates, which allows task-related information to be included in predetermined expressions [1, 2]. This helps make the system scalable within the context of the project while simultaneously avoiding the robots' behaviours becoming unpredictable. Furthermore, given that the interaction designed for the L2TOR project involves a child in a language tutoring setting, it is especially important that this spoken language is as close to natural-sounding as possible, as children have been shown to pay attention to non-verbal aspects of speech when acquiring a language (see e.g., [3]).

The project proposal specifies the output module for the number domain along three tasks that, in essence, focus on the design, development, and implementation of a natural language generation (NLG) system for multimodal and multilingual output. As specified in Deliverable 1.1, the lessons for the number (or math) domain concern not only number words (one, two, three, etc.) but also language regarding (pre-)mathematical concepts, such as weight, size, and quantities. Target words will be taught through a range of game-like activities to be played with the robot and the tablet application (note that the math lessons have been altered to some extent since Deliverable 1.1 was submitted). Activities include the counting of animals in a zoo scenario (introduction to quantities), making of a bouquet in a flower shop (larger quantities), baking of breads and cakes in a bakery (adjectives), finding of animals in a zoo (comparatives and superlatives), and feeding of animals at a farm (large quantities). Although these activities are centred around the number domain, we will be able to adjust parts of the implementation (e.g., the manner in which feedback is generated) to fit the contents of the other domains.

Considering expectations regarding the general implementation of the system, the robot is expected to be able to not only generate spoken language output for multiple languages and according to various levels of proficiency, but to also facilitate appropriate non-verbal behaviours, such as prosodic cues, gaze, and gestures. While the project builds on existing work, e.g., [4, 5, 6], these need to be adapted to the specific requirements of the lesson series, and to specific constraints of SoftBank Robotics' humanoid NAO robot (and its possibilities for multimodal output). Some of the challenges faced so far include the (in the use context) unnatural pronunciation of certain words by the robot, for which we have come to rely on hand-crafted phonetic representations, creating syntactic templates to support multilingual utterances for which content is derived from an interpretation of user actions, and developing gestures that are meaningful and understandable when produced by the NAO robot. In addition, from the perspective of NLG, it should be noted that much of the curriculum has been specified and will follow some predetermined flow, resulting in less flexibility regarding intelligent output generation.

To summarize, building on previous research, the contribution of WP 6 to deliver on the promise of a multilingual and multimodal social robot suited to teach young children a second language through playful interactions involves the development of an output generation module capable of – from high-level semantic information provided by other modules – generating natural language in the context of and befitting the content of a series of largely predetermined tutoring sessions. When spoken by the robot, generated utterances are produced as natural-sounding, child-directed synthesized speech, addressing non-verbal aspects such as prosody, and accompanied by non-verbal behaviours, such as gaze and (co-speech) gestures, when appropriate. This output generation module is to be implemented in NAO robot as part of an integrated language tutoring system.

The following sections will explain in more detail the progress made towards this goal for the number domain from the perspective of WP 6.

2 System Architecture

To illustrate the architecture of the integrated system at the time of writing, Figure 1 provides a graphical overview of how the different modules are communicating with each other. Note that all communication between managers runs through a broker, the ConnectionManager (not shown in Figure 1, instead see Deliverable 3.1), which was developed as part of WP 3 and is described in Deliverable 3.1.

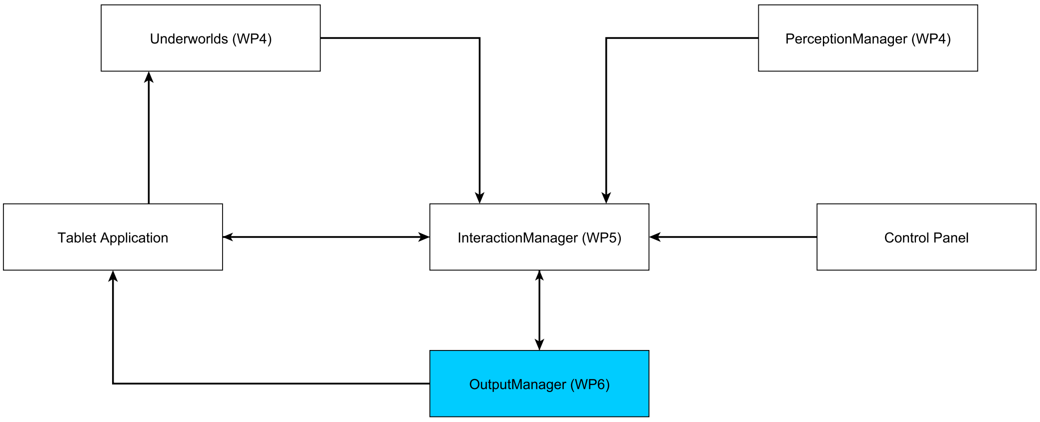


Figure 1: The integrated technical tutoring system and the role of the OutputManager.

2.1 Output Module

The current architecture of the OutputManager is shown in Figure 2. The main entry point of the module is also called OutputManager. It takes care of all communication with the other modules within the integrated system.

Each session from the lesson series has its own scenario, of which the output is stored in a JSON file. This allows the InteractionManager to request the OutputManager to present a certain task of the current scenario, and the OutputManager will be able to read from this JSON file all the output it needs to produce for that particular task. To be able to generate correct referents to objects on the tablet, we also provide a dictionary with the correct translations in all possible L1 or L2 languages.

The submodules take care of specific output needs:

- TabletManager performs speech output from the tablet speakers, using a set of pre-recorded utterances from a native speaker;
- OutputRealizer actually sends output to the NAO by calling upon its Text-To-Speech (TTS) engine and other behaviours (gestures, gaze, etc.). It is able to call back to the main OutputManager module, for example, to indicate that output has completed;

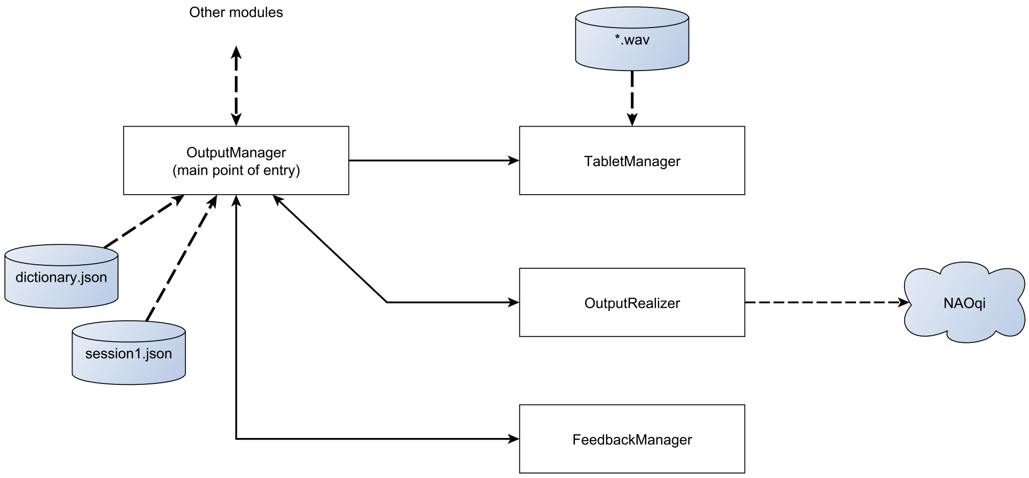


Figure 2: The internal architecture of the Output Module.

- FeedbackManager uses syntactic templates to build complex feedback utterances, which are then returned to OutputManager so that they can be send to the OutputRealizer (see Subsection 4.2).

This architecture is designed with the underlying idea that each submodule should have its own purpose and should be (relatively) independent from the others. For example, if we were to do an experiment with a virtual avatar instead of the NAO, we would only have to replace the OutputRealizer submodule as this is the only part of the output generation that actually sends output to the robot.

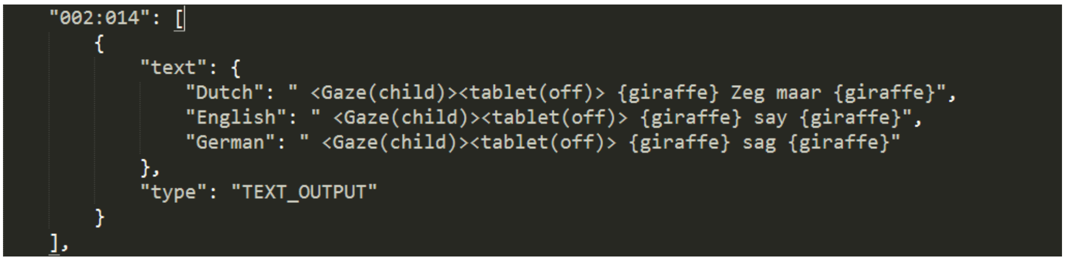
3 Tasks of the Output Module

The OutputManager is able to perform the functions listed below within the integrated tutoring system. It only receives input, in the form of function calls, from the InteractionManager, because that is the module managing the flow of the sessions. There are also certain variables, such as the intended pair of L1 and L2 and the name of the child, that will be send by the InteractionManager at the beginning of the session and then stored within the OutputManager for later use.

Output is directly delivered to the child in the form of verbal and non-verbal behaviour of the robot and speech from the tablet. Furthermore, the InteractionManager is notified when output has completed, and the OutputManager is able to call functions on the tablet, for example, to dim the screen. The following subsections will describe all the functionalities that the OutputManager provides to the system in more detail.

3.1 Load Session

The evaluation study will contain multiple sessions with each child, spaced out over a number of weeks. There is a memory contained within the InteractionManager that keeps track of the progress of a child, so that when a new session takes place the InteractionManager will know what content to load, and send this *load session* command to the OutputManager. The OutputManager has a collection of JSON files, that each describe all of the content belonging to a particular session (see Figure 3). These files are automatically generated from the scenario descriptions that were designed in Excel as a part of Deliverable 2.1.



```
"002:014": [
    {
        "text": {
            "Dutch": " <Gaze(child)><tablet(off)> {giraffe} Zeg maar {giraffe}",
            "English": " <Gaze(child)><tablet(off)> {giraffe} say {giraffe}",
            "German": " <Gaze(child)><tablet(off)> {giraffe} sag {giraffe}"
        },
        "type": "TEXT_OUTPUT"
    }
],
```

Figure 3: Example of session information stored in the JSON file.

3.2 Give Task

This function presents a task to the child. Each session with the child is made up of a number of consecutive tasks. A task is always associated with a way to complete it, in order to progress to the next task of the session. In the case of an introductory task, this trigger could be as simple as the robot finishing its output. It could also require some action from the child, such as repeating a word or touching an object on the tablet screen. The InteractionManager is responsible for checking whether the specific criteria for moving to the next task have been met. If this is the case, the InteractionManager will decide to move to the next task and again call upon the OutputManager to present this task to the child.

The output belonging to a task can consist of verbal and non-verbal output from the robot, as well as sound files played from the tablet. Furthermore, to help direct the attention of the child we are able to turn off (black out) the tablet screen and turn it on again at specific points of an utterance, so that when the robot requests the child to repeat a word (an action that does not involve the tablet), the tablet will turn off to avoid distractions. This is an example of a combined output command for the robot:

> Now, I think there's a very important task for us! <tablet(on)>The monkey is loose and we have to put it in its cage! Put <pointAt(tablet)><Gaze(tablet)>the {Affe} in its cage

This example includes an utterance in both L1 and L2, where the switch to L2 is indicated with curly brackets. Furthermore, interaction with the tablet (turning on the screen), a robot gesture (pointing) and a change in the gaze direction of the robot are timed to occur at specific points during speech. The OutputManager runs through this command, transforms tags to be compatible with the TTS engine in question, and makes corrections with respect to the pronunciation where needed (see Subsection 4.4).

3.3 Give Feedback

If a task requires any input from the child (in the form of speech activation or an interaction with the tablet), and this input is received, the InteractionManager will prompt the OutputManager to give feedback. This feedback is formed by taking templates, based on feedback strategies that were investigated in WP 2, and filling the gaps with variables that are specific to the current task. This is described in more detail in Subsection 4.1.

3.4 Give Break

From the experiment described in [7], we observed that engagement levels of the participating children tend to drop towards the end of a session of about 15–20 minutes. Although there will be more variation in content for the full L2TOR evaluation study, because the session length is similar, there is a risk that children will experience the same drop in engagement. To counteract this, we are considering to implement breaks into the flow of a session, where the child and robot perform an unrelated task such as dancing, stretching, or a simple game that is unrelated to the learning experience (see e.g., [8]). After running pilot studies without breaks, a decision will be made whether scheduled breaks will be needed for the system.

3.5 Resume Interaction

After an interruption of the session, for example, in the case of a scheduled break (as discussed in the previous section) or if the interaction had to be paused by an experimenter, for any reason, from the control panel, there should be a way for the child to be reintroduced into the flow of the session. For example, this means that the robot could welcome the child back into the room after a restroom break, or it could announce its intention to continue working on the main tasks at hand after a stretching exercise, after which the interrupted task could be repeated or a next task could be presented.

3.6 Request Answer

Every task has a certain criterion that needs to be met, in order to trigger the next task. This could be as simple as the robot completing its output, or it could also involve input from the child in the form of speech activation or an interaction with the tablet. If this input is indeed expected, but the child appears unresponsive, the InteractionManager will ask the OutputManager to request an answer from the child. This is based on how a human tutor would respond in a similar situation, thereby showing that the robot is ‘aware’ that it is the turn of the child to perform an action in order to proceed.

Currently an answer is requested based on the criterion of the current task, including: type of task with the objects and spatial relations that it might require. Based on a template for each type of criterion (speech activation, tablet interactions) a sentence is constructed that, in a very concise way, reminds the child of what is expected in order to proceed:

> Can you put the monkey into the cage?
>
> Can you put *obj_1 spatial_relation obj_2*?

Building this sentence is not straightforward because variables such as obj_1 can be in either L1 or L2, and in some cases there can be several instances of obj_1 or obj_2. If the task was first introduced within the normal flow of the script as “Can you put the *smallest* monkey into the cage *with most animals*”, upon requesting an answer the same descriptor should remain. Currently, the OutputManager has no knowledge of context regarding obj_1 and obj_2, therefore we are exploring the use of a discourse model to keep track of all objects in the scene and their discriminating features. Sometimes, *finding* the target object is actually part of the task, as in the following example:

> Can you touch the area with most animals?

The system expects the child to touch the lake in order to finish the task, so a naive approach to requesting an answer would yield: “Can you touch the lake?”, effectively giving away the answer.

3.7 Offer Help

After a task that requires a response from the child was presented, but not answered after a certain amount of time, the InteractionManager will try to request an answer (as shown above). If there is then still no response from the child, we assume that further guidance is needed. To avoid adding a lot of extra time to the duration of the session, the robot will choose to help the child by performing the desired action for them. For example, when one object has to be placed inside of another object, the robot will perform a gesture that makes it seem like it is interacting with the tablet, while the object simultaneously moves to the correct location on the screen. This will allow the session to be resumed while simultaneously mitigating the interruption in the flow of the interaction.

3.8 Grab Attention

WPs 4 (input) and 5 (InteractionManager) are working on the implementation of a feature to estimate and track the affective state of the child during a tutoring session. If the engagement starts to drop, we plan to design an attention grabber to draw the child back into the interaction. The ability to request an answer performs a similar role, but it is specific to the task at hand. The attention grabber is a generic utterance used when the attention seems to drop for several consecutive tasks based on the teacher-child observations as described in Deliverable 1.2.

3.9 Interrupt Output

It is possible, especially in the case of requesting an answer to a task, that the child actually performs an action while the robot is speaking. To make the robot appear as an intelligent conversational partner, it should be able to notice these changes in context and act accordingly. This means that it should stop talking when it is no longer relevant to request an action from the child, and instead provide feedback with respect to the action taken by the child.

3.10 Output Completed

This function is currently the only response that flows back from the OutputManager to the InteractionManager. It triggers once the output from the robot has completed. This ensures that the InteractionManager knows when to start checking for task completion criteria, and when to start requesting an answer if the child does not respond. The disadvantage is that objects on the tablet do not become enabled until the robot finishes all of its output (verbal and non-verbal), which makes the tablet unresponsive and slows down the interaction. We are, therefore, experimenting with moving this signal forward in the robot output, so that the signal is sent slightly before the output actually finishes, so that children can start moving objects around on the tablet right as the robot finishes speaking.

3.11 Logging

All modules of the integrated system, including the OutputManager, implement logging for debugging purposes, and so that tutoring sessions can be thoroughly analysed afterwards. Log files are stored on the tablet device, with one file per module per session.

We are currently logging every OutputManager function call, as well as all output commands that get sent to the robot (verbal and non-verbal). Because functionalities such as requesting an answer and giving feedback can introduce variation between the number of times each child was exposed to a target word in L2, it is also important to keep track of the specific output, and whether this output was L1 or L2.

4 Design of Robot Behaviour

4.1 Feedback

The FeedbackManager is one of the sub-modules of the OutputManager and, as the name suggests, handles the feedback to be provided in response to the child's (un)successful execution of a task. Ideally we would learn over the course of the interaction(s) the feedback strategy with the biggest pay-off for the child. As [9] showed (and as explained in Deliverable 2.1), the individual differences between children are larger than the effect of different types of feedback, which emphasizes the importance of personalisation for every child. Some children respond better to explicit feedback than implicit feedback or become more engaged with positive feedback. However, the level of adaptation and personalization, as is the case in general with respect to the experimental evaluation of the project, should be kept to a minimum for reasons of experimental consistency (as mentioned in Deliverable 2.1).

Currently, feedback is generated following a syntactic template to fit a specific task (cf., [10]) where the gaps in the template are filled by information regarding objects in the scene presented on the tablet with which the child interacts. The information necessary to build the expression varies per task. This may include information regarding the objects with which the child was expected to interact, the objects with which the child has interacted, the relation between the objects, the language used to refer to the objects, and whether or not the task has been successfully completed by the child. For example, for an object movement task the child is asked to put one object, e.g., a giraffe, inside another object, e.g., a cage. When the child does not manage to do this correctly, the following syntactic template may be used (when L1 = Dutch):

> *”Nee dat klopt niet.* < `article` > < `target_word` > *moet* < `relationship_word` > < `helper_article` > < `helper_word` >. *Probeer het nog maar een keer.”*

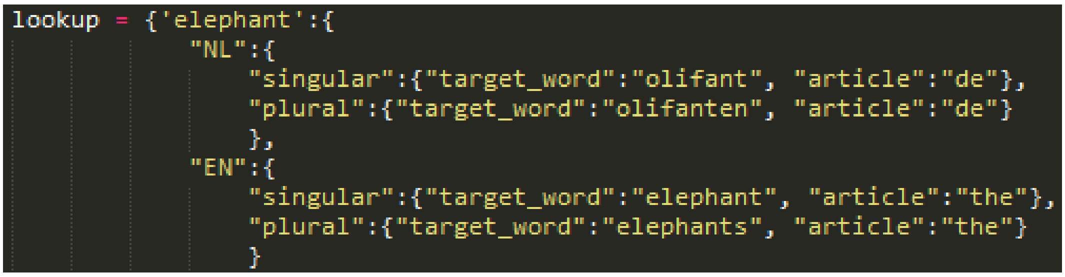
Here, the child received negative feedback as well as some comment with respect to the specifics of the task to be completed. The gaps in this template are given by `article`, `target_word`, `relationship_word`, `helper_article`, and `helper_word`, where, for this task

- `article` = **the** (English article, belongs to target word)
- `target_word` = **giraffe** (animal, English target word)
- `relationship_word` = **in** (spatial relation between target and helper)
- `helper_article` = **de** (Dutch article, belongs to helper word)
- `helper_word` = **kooi** (Dutch word for *cage*)

This results in the utterance

> *”Nee dat klopt niet. The giraffe moet in de kooi. Probeer het nog maar een keer.”*.

These templates may be deceptively simple; in actuality they require a great deal of information regarding the current scene and user actions to be interpreted by the various modules and forwarded to the output module, to then be interpreted and used by the FeedbackManager. This difficulty has been discussed previously in relation to the task of requesting an answer (see Subsection 3.6), for which we are considering the use of a discourse model. Moreover, for the retrieval of task-specific information necessary for the construction of feedback utterances we will rely on a lookup table which holds all information relevant to the objects in play, for example, such as listed for the example given. Figure 4 exemplifies how this information may be stored for the word *elephant*.



```
lookup = {'elephant':{
            "NL":{
                "singular":{"target_word":"olifant", "article":"de"},
                "plural":{"target_word":"olifanten", "article":"de"}
                },
            "EN":{
                "singular":{"target_word":"elephant", "article":"the"},
                "plural":{"target_word":"elephants", "article":"the"}
                }
```

Figure 4: Example of how information may be stored in the lookup table.

4.2 Gestures

One of the main advantages of the NAO robot as a tutor is its humanoid appearance as well as its physical presence in the real world. This allows us to make use of human-like, non-verbal behaviours, such as gestures, to support the interaction between robot and child and facilitate learning in a natural way (e.g., to aid word learning or as a way of grabbing or holding attention).

To investigate the effect of (iconic) gestures on L2 word learning, we conducted a study in collaboration with Bielefeld University (WP 5), in which children, all native speakers of Dutch, were taught the English names of animals in a playful interaction based on the children's game *I spy with my little eye*, which was designed specifically for the setup as proposed for the L2TOR project, using the robot and a tablet interface. The experiment was adapted from [11], a previously conducted study in the context of WP 5 which set out to investigate the use of an adaptive tutoring approach based on Bayesian knowledge tracing to aid the acquisition of L2 vocabularies. Participants in the [11] study, however, were adults. We adjusted the experiment so that we could examine the effects of the robot's use of iconic gestures on learning in addition to the effects of adaptive language tutoring, as well as make the procedure more appropriate for children. To clarify what is meant by iconic gestures, simply put, iconic gestures are movements that can said to be iconic for the linguistic unit(s) they may co-occur with, as these movements show a clear semantic match with the information communicated [12].

During the experiment, participants were presented with pictures of the animals, of which one target and others distractors. The robot would ask the child to select the correct picture belonging to the target word, the English name of the animal depicted. This target word was uttered either with or without an iconic gesture representing the animal in question. See [7] for more details regarding the design of this experiment (see Appendix).

As the use of gestures has been shown to be an effective scaffolding technique for learning novel L2 words (e.g, [13, 14, 15, 16, 17]), we expected a higher learning gain when L2 words were presented with as opposed to without congruent iconic gestures. Moreover, in line with findings of [18], we expected gestures to reduce knowledge decay over time, meaning participants were expected to retain a higher number of L2 words when these words had been presented with congruent iconic gestures during training. To test our assumptions, prior to the tutoring interaction with the robot, participants were tested on their prior knowledge with respect to the target words that were part of the experiment. Using this indication of prior knowledge and by conducting post-tests similar to the pre-test, we could calculate scores representing the participant's learning gain immediately after their interaction with the robot as well as seven days (minimum) after their participation in the experiment (retention). Preliminary results show that, although, on average, all participants regardless of assigned condition performed better on the immediate post-test than on the pre-test, children who had been assigned to the gesture condition did not score significantly higher than those who had not been exposed to the gestures. However, on the retention test we did find a significant difference, as the children indeed retained a higher number of L2 words when they had been presented with iconic gestures. Although we did not find a significant effect of gestures on learning gain immediately following the children's interaction with the robot, results from the retention test do show the added value of the use of gestures during vocabulary training. The full results of this experiment will be presented in a paper currently in preparation.

On a side note, an additional finding concerns the pronunciation of the English words. We found that the Dutch word for horse, *paard*, was often confused with the English word *bird*, as children would frequently select the horse when prompted with *bird*. This was especially apparent on the post-tests, as the children would often repeat the target word they had just heard, leading them to say *paard* when they were prompted with *bird* and, thus, often mistaking the English *bird* for the Dutch *paard*, even when the child had not mistaken the two words during training. Although this finding emphasizes the importance of natural-sounding speech synthesis, interestingly, the target words for the pre- and post-tests were recorded by a native speaker of English.

At present, gestures, such as the iconic gestures used in the experiment described in [7], have been manually implemented using SoftBank's Choregraphe Suite software, a package which supports visual programming of the robot's behaviours. By manipulating the posture of either the virtual or the physical robot, a technique called puppeteering, we stored specific poses of the robot in key frames, which, when acted out by the robot, resulted in animations resembling gestures iconic for the animals used in the experiment. See Figure 5 for an impression of the workspace. This approach, however, is rather labour-intensive. Ideally, we would generate these gestures automatically (see Section 6).

4.3 Gaze

The gaze direction is described in the scripts of each session (see Deliverable 2.1). To create gaze behaviour that is more robust to changes in the positioning of robot and child, we use face tracking to locate the actual position of the child, rather than using fixed positions of where we expect the child to be sitting.

We extended the existing face tracking functionality of the NAO robot to store the last known position of the child. Because the robot is constantly shifting focus from child to tablet, having to move its head and thereby losing track of the child's position, remembering the last known head position (joint angle) of when the child was still visible helps to quickly finding the child again.

4.4 Speech Synthesis

The NAO robot comes with default TTS engines. For English, the engine is powered by Nuance and for the other languages it is Acapela. The voice itself has a character that matches well with the appearance

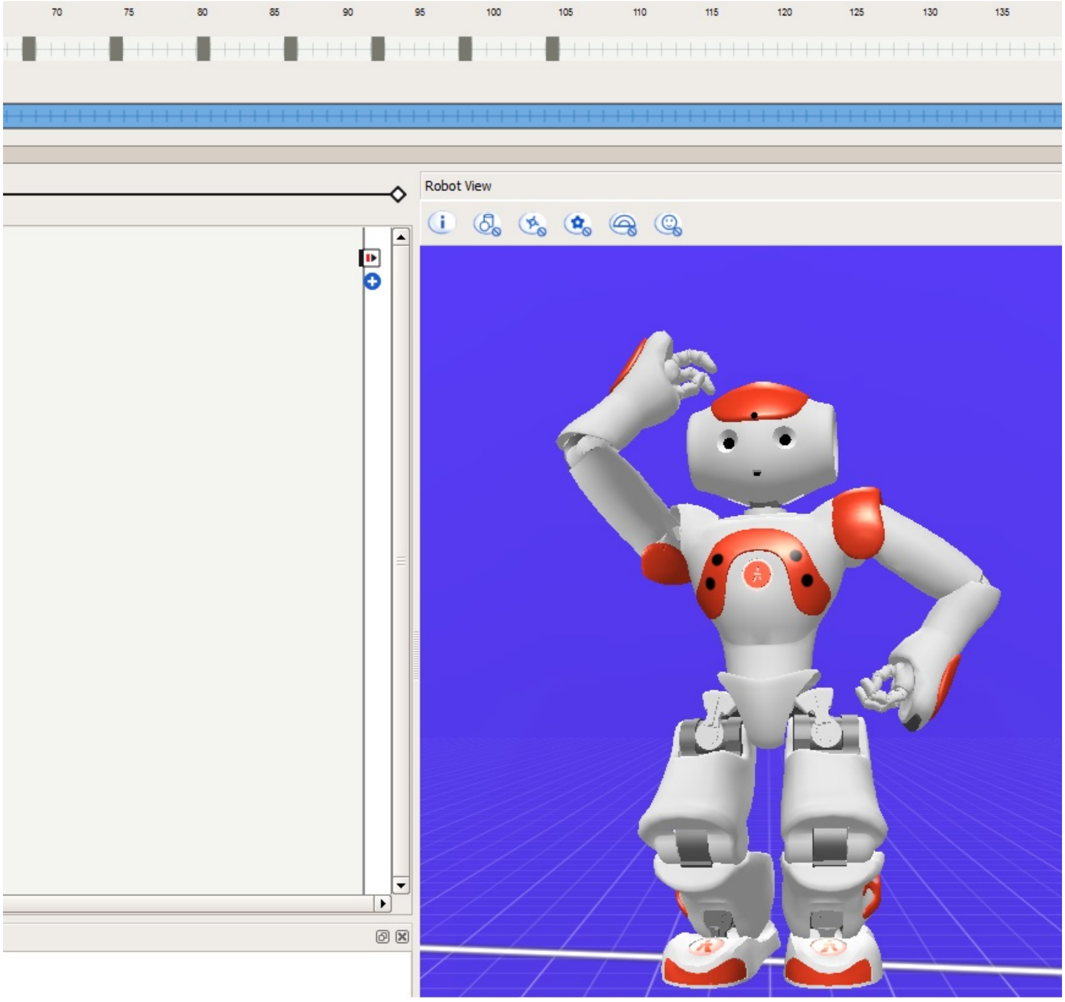


Figure 5: Part of the Choregraphe workspace. At the top, key frames are visible that represent the robot's pose at that point in time. The example gesture shown by the virtual robot is the iconic gesture for *monkey* used for the experiment described in [7].

of the robot (similar to that of a child). Furthermore, it is possible to change pitch, speed and volume of the voice.

The voices do have a different sound between languages, where some pairs are more closely related (Dutch and English) than others (German and English). However, we could not find alternative solutions where the same voice actor was used as input for multiple languages, to provide a consistent sound while switching languages. In our first experiments where the robot combined Dutch and English, children did not seem to notice this difference so we have for now decided to push forward with the built-in voices of the NAO robot.

There were several other considerations regarding the speech output of the robot, namely that it should allow:

- Non-verbal behaviour (gestures, gaze shifts) that is timed with specific locations within an utterance (as specified in the storyboards, see Deliverable 2.1);
- Interruptions, to give a sense of intelligence through awareness of context;
- Fast switching back and forth between two languages;
- Triggering actions within other parts of the system, timed with locations within an utterance (e.g. dimming the screen when the tablet is irrelevant).

To correct words that are mispronounced by the robot, we will create a dictionary that can be easily maintained by all partners while translating the scenarios for tutoring sessions. All entries in this dictionary that are found in the sentence to be send to the TTS engine will then be replaced by their corrected versions. Both Nuance and Acapela accept a phonetic representation of a word. To exemplify the use of these phonetic transcriptions, when referring to the tablet when the L1 is Dutch, the TTS will pronounce the word *tablet* as if it were to mean a pill of sorts (e.g., medicine) rather than a tablet computer. By using the following phonetic representation of the word

t E: b l @ t

we are able to use the pronunciation that more closely resembles the manner in which a tablet computer is commonly referred to (Anglicism) in Dutch.

5 Tablet Application

5.1 Sound

Because the quality of pronunciation by the robot will still not approach that of a native speaker, the first exposure to a new target word in the L2 will be done by playing a recording of a native speaker pronouncing the target word from the tablet. In addition, this increases the different exposures of new target words which has been shown to support L2 learning (see Deliverable 1.1).

5.2 Coordinating Actions

Actions such as turning off the tablet to guide attention, or helping the child by showing how to perform an action, need to be timed correctly between the robot and the tablet. For example, if the robot is moving its hand to move an object on the tablet, after first introducing this act by saying “Here, let me show you..”, the timing of the hand gestures should coincide with the animation on the tablet screen to make the act seem realistic.

The NAOqi interface and its ALAnimatedSpeech interface allow for events to trigger at specific points within a speech output, enabling us to activate not only behaviours on the NAO itself but also any arbitrary function within our module. For example, the command:

> Here, let me show you.. $begin_tablet_help(elephant_1, cage_1) ^start(moving/tablet)

will trigger an event ”begin_tablet_help” after the robot finishes speaking, which can prompt the OutputManager to send a message to the Tablet Application that it is time to animate moving the particular object while at the same time the robot will start its gesture of pretending to move something on the tablet.

6 Conclusion and Future Directions

This deliverable discussed the progress made towards the realization of the output module (WP 6) of the integrated tutoring system. Although the focus has been on the development of an output module for the number domain, many of the functionalities described will carry over or will be adapted for the other domains. With respect to NLG, we are bound by the scripted scenarios (as described in Deliverables 1.1 and 2.1), which has led us to use a template-based approach when the dynamic generation of language is required. Other challenges we have encountered involve the robot's non-verbal behaviour, such as the development of appropriate gestures (currently a manual effort) and the tuning of the speech synthesis (e.g., correcting of pronunciation errors through the phonetic transcription of utterances), which have proven to be laborious, time-consuming efforts. Moreover, the tuning of the speech synthesis and construction of appropriate syntactic templates has been especially challenging because of the multilingual nature of the project.

Future work includes further investigating and identifying the added value of gestures for children learning a second language. To speed up the process of generating gestures and to make them more human-like and 'spontaneous', we intend to use the Kinect to teach the robot gestures by demonstration (e.g., as in [19]). We plan to conduct an experiment to see how well gestures, recorded from a human, map onto a humanoid robot and how much of the meaning is lost due to the physical limitations of the robot.

Furthermore, we are exploring ways in which the generation of language (e.g., output with respect to answer requests as well as feedback) can be made more context-aware, through the use of a discourse model. This will make it possible to provide the children with output that is specifically generated for the current task or their actions taken.

References

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

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

[3] Peter F. Dominey and Christelle Dodane. Indeterminacy in language acquisition: the role of child directed speech and joint attention. *Neurolinguistics*, 17(23):121–145, 2004.

[4] Kirsten Bergmann and Stefan Kopp. *GNetIc – Using Bayesian Decision Networks for Iconic Gesture Generation*, pages 76–89. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[5] Kirsten Bergmann and Stefan Kopp. Modeling the Production of Co-Verbal Iconic Gestures by Learning Bayesian Decision Networks. *Applied Artificial Intelligence*, 24(6):530–551, 2010.

[6] Ielka van der Sluis and Emiel Krahmer. Generating Multimodal References. *Discourse Processes*, 44(3):145–174, 2007.

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

[8] Aditi Ramachandran, Chien-Ming Huang, and Brian Scassellati. Give Me a Break!: Personalized Timing Strategies to Promote Learning in Robot-Child Tutoring. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pages 146–155, New York, NY, USA, 2017. ACM.

[9] Mirjam de Haas, Peta Baxter, Chiara de Jong, Emiel Krahmer, and Paul Vogt. Exploring Different Types of Feedback in Preschooler and Robot Interaction. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pages 127–128, New York, NY, USA, 2017. ACM.

[10] Kees Van Deemter, Emiel Krahmer, and Mariët Theune. Real Versus Template-Based Natural Language Generation: A False Opposition? *Computational Linguistics*, 31(1):15–24, March 2005.

[11] Thorsten Schodde, Kirsten Bergmann, and Stefan Kopp. Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pages 128–136, New York, NY, USA, 2017. ACM.

[12] David McNeill. So you think gestures are nonverbal? *Psychological Review*, 92(3):350–371, 1985.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Susan Wagner Cook, Zachary Mitchell, and Susan Goldin-Meadow. Gesturing makes learning last. *Cognition*, 106(2):1047–1058, 2008.

[19] Heni Ben Amor, David Vogt, Marco Ewerton, Erik Berger, Bernhard Jung, and Jan Peters. Learning responsive robot behavior by imitation. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3257–3264. IEEE, 2013.

Exploring the Effect of Gestures and Adaptive Tutoring on Children's Comprehension of L2 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

The L2TOR project explores the use of social robots for second language tutoring. This paper presents an experiment in preparation to investigate the effects of two educational scaffolding features (adaptation/personalization and iconic gestures), when used by a robot tutor, on children's comprehension of animal names in a foreign language. Participants will be children between the ages of four and five. The study is scheduled to take place in March 2017.

CCS CONCEPTS

•Computing methodologies → Cognitive robotics; Probabilistic reasoning; •Applied computing → Interactive learning environments; •Human-centered computing → *Empirical studies in HCI;*

KEYWORDS

Language tutoring; Assistive robotics; Education; Bayesian knowledge tracing; Human-robot interaction

ACM Reference format:
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 ACM HRI conference, Vienna, Austria, March 2017 (R4L workshop at HRI 2017),* 6 pages.
DOI: 10.475/123_4

1 INTRODUCTION

The L2TOR project aims to design and develop a robot tutor capable of supporting children of four to five years old in the acquisition of a second language by interacting naturally with them in their social and referential environment through one-to-one tutoring interactions [1]. The robot used for the L2TOR project is the SoftBank Robotics NAO humanoid robot. The NAO robot is capable of speaking multiple languages, readily able to switch between them, which provides the possibility to vary the amount of the child's native language (L1) and the second language (L2) to be taught. Furthermore, the physical presence of a robot is shown to improve learning gains compared to its two-dimensional counterparts (e.g. Leyzberg et al. [12]).

This three-year project will result in an integrated lesson plan, which is expected to contain 24 lessons spanning three different domains (math, space, and mental state). To design these lessons, we analyze the way human tutors interact with children and investigate how different functionalities of the robot can be used to ensure a natural and productive interaction. In this paper, we propose an experiment to evaluate two such functionalities: personalized lessons by adjustment of the level of difficulty of the subject matter to the level of proficiency of the learner and the use of gestures when introducing the L2 words. We expect that both concepts will help to create and maintain common ground with the child, while also increasing comprehension and memorization potential of new words in the L2.

The importance of personalized adjustments in the robot's behavior has been substantiated in recent research showing that participants who received personalized lessons from a robot (based on heuristic skill assessment) outperformed others who received a non-personalized training [12]. Suboptimal robot behavior (e.g. distracting, incongruent or in other ways inappropriate social behavior) can even hamper learning [10].

One of the main advantages of choosing a humanoid robot as a tutor is its physical presence in the world, allowing for interactions similar to those between humans. Because of its anthropomorphic appearance, we tend to expect human-like communicative behavior

R4L workshop at HRI 2017, Vienna, Austria

DOI: 10.475/123_4

[*] Tilburg center for Cognition and Communication

[‖] Cluster of Excellence Cognitive Interaction Technology

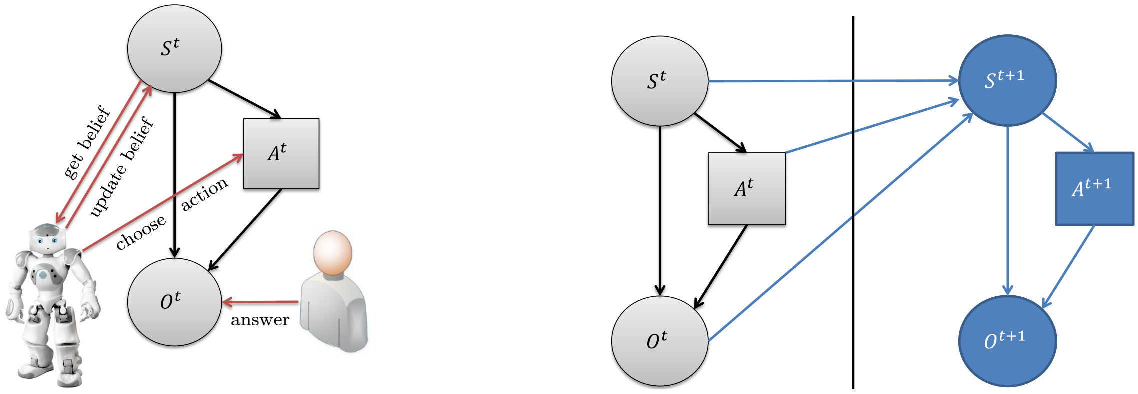


Figure 1: Dynamic Bayesian Network for BKT: With the current skill-belief the robot chooses the next skill S^t and action A^t for time step t (left). After observing an answer O^t from the learner, this observation together with action A^t and the previous skill-belief S^t are used to update the skill-belief S^{t+1} at time $t + 1$ (right) [18].

from the robot, including proper use of non-verbal communication. Robots that perform gestures are perceived in a more positive way than those that use only speech [16].

In Section 2 we explain our previous work to evaluate adaptive learning, which is used as a starting point for the experiment described in this paper. We then introduce iconic gestures and describe how they could be used to increase learning gain in a human-robot tutoring context in Section 3, followed by our main research questions in Section 4. Section 5 outlines the design of the proposed experiment. We intend to start data collection in March 2017.

2 PREVIOUS WORK

2.1 Adaptive language tutoring with a robot

In previous work we developed a novel approach to personalize language tutoring in human-robot interaction [18]. This adaptive tutoring is enabled through a model of how tutors mentalize about learners – by keeping track of their knowledge state and by selecting the next tutoring actions based on their likely effects on the learner. This is realized via an extended model that combines knowledge tracing (of what the learner learned) with tutoring actions (of the tutor) in one causal probabilistic model. This allows for selecting skills and actions based on notions of optimality – here the desired learner's knowledge state as well as optimal task difficulty – to achieve this for a given skill.

The approach is based on Bayesian Knowledge Tracing (BKT) [4], a specific type of Dynamic Bayesian Networks (DBNs). The model consists of two types of variables, namely the *latent variables* representing the belief state of 'skills' to be acquired (e.g. whether a word has been learned or not) and the *observed variables* representing the observable information of the learning interaction (e.g. whether an answer was correct or not). In our proposed model, each latent variable can attain six discrete values, corresponding to six bins for the belief state (0%, 20%, 40%, 60%, 80%, 100%) representing whether a skill is mastered or not as a discretized probability distribution. That is, we reduce the complexity we would get through continuous latent variables but also attain more flexibility. The observed variables remain binary and still represent whether a learner's response is correct or not (see Figure 1). Moreover, the following update of the belief state of the skill, i.e. the skill-belief, at time $t + 1$ is not only based on the previous skill-belief, but also on the chosen action and the previous observation at time t.

Based on this model, two types of decisions are made, (1) which skill would be the best to address next, and (2) the choice of action to address that skill. Regarding the former, we employ a heuristic maximizing the beliefs of all skills while balancing the single skill-beliefs among each other. This strategy is comparable to the vocabulary learning technique of *spaced repetition* as implemented, for instance, in the Leitner system [11]. Regarding the choice of action, the model enables the simulation of the impact each action has on a particular skill. To keep the model simple, the action space of the model consists of three different task difficulties (easy, medium, hard). Consider an example where the skill-belief appears relatively high, such that the skill is nearly mastered by the learner. In this case, a less challenging task would only result in a relatively minor benefit for the training of that skill. In contrast, if we assume the skill-belief to be rather low and a very difficult task is given, the student would barely be able to solve the task, likewise resulting in a smaller (or non-existent) learning gain. Instead, a task of adequate difficulty, not needlessly simple nor too complicated for the student to solve, will result in a higher learning gain [5]. This helps to position the robot as a capable instructor that uses these scaffolding techniques to help children acquire new skills beyond what they could have learned without help, by bringing them into the zone of proximal development (ZPD) [22].

2.2 Evaluation

When implemented in the robot language tutor, the model will enable the robot tutor to trace the learner's knowledge with respect to the words to be learned, to decide which skill (word) to teach next, and how to address the learning of this skill in a game-like tutoring interaction. For the experiment as described in [18], participants were asked to learn ten vocabulary items German – 'Vimmi'

(Vimmi is an artificial language that was developed to avoid associations with other known words or languages for language-related experiments [13]). The items included colors, shapes and the words 'big' and 'small'. During the game, the robot would introduce one of the Vimmi words. A tablet then displayed several images, one of which satisfied the Vimmi description (e.g. one object that is blue) and a number of distractors. The participant was then asked to select the image corresponding to the described item. Participants learned vocabulary items in one of two conditions, either in the condition with the adaptive model or in a non-adaptive (random) control condition. In the adaptive condition, the skill to be taught and the action to address the skill were chosen by the model as described above. Participants' performance was assessed with two measures: (1) learners' response behavior was tracked over the course of the training to investigate the progress of learning, and (2) a post-test was conducted on the taught vocabulary in the form of both L1-to-L2 translations and L2-to-L1 translations to assess participants' state of knowledge following the intervention.

Analysis of participants' response behavior over the course of training indicated that the participants learned the L2 words during the human-robot interaction (see [18] for more detailed results). Importantly, they learned more successfully with our adaptive model as compared to a randomized training. That is, the repeated trials addressing still unknown items as chosen by the adaptive model (until the belief state about these words equaled that of known items) outperformed the tutoring of the same material (same number of trials and items) but in randomized order. In the post-test, however, there was no significant difference across experimental conditions, despite a trend towards increased performance in the adaptive model conditions as compared to the controls.

3 ICONIC GESTURES

A growing body of evidence suggests that iconic gestures bear a great potential to enhance learners' memory performance for novel L2 words. Iconic gestures are movements that have a formal relation (in form or manner of execution) to the semantic content of the linguistic unit they describe [14]. In other words, the gesture elicits a mental image that relates strongly to the word or words that it links to. As an example, the word *bird* could be described by an iconic movement of stretching both arms sideways and moving them up and down, symbolizing the flapping of wings. The supporting effect of iconic gestures on L2 vocabulary learning by providing a congruent link between the words to be learned and gesture being observed or imitated has been shown in various studies (e.g. [6, 9, 13, 15, 19]). A recent overview of how gestures contribute to foreign language learning and possible explanations for this effect is given by Hald et al. [8]. Although they focus mainly on students *performing* or re-enacting the gestures, merely observing a gesture is shown to aid learning as well. Research conducted by Tellier [19] and De Nooijer et al. [6] investigated the role of gestures with respect to children and word learning. The effect of gestures is shown to depend on the students' gender, language background and existing experience with the L1 [15].

When considering the use of an artificial embodied agent as a tutor, the positive effects of gesturing seem to apply as well, as shown by Bergmann and Macedonia for a virtual tutor [2], and by Van Dijk et al. for a robotic tutor [20]. An additional benefit of implementing non-verbal behavior is to improve the way the robot is perceived, making it seem more human-like[17]. The challenge of mapping non-verbal behavior to the robot lies in the fact that each act needs to be carefully designed and choreographed to coincide with the corresponding word or sentence. There are limits to the degrees of freedom, the working space (i.e. the physical reach) and smoothness of motion that the robot has to offer. As an example, Figure 2 shows an attempt at making an iconic gesture for *rabbit*. The right arm has to take an unnatural position, which may result in an uncanny feeling for the observer. The NAO robot also has only three fingers and they cannot move independently, therefore finger-counting and similar subtle motions do not transfer to the robot without modification. The challenge lies in finding ways to work around these limitations, while still taking advantage of the added value of non-verbal communication. The gestures that were designed for this experiment have been exaggerated beyond what the human alternatives would look like. For example, when imitating a monkey the robot will bend its knees and shift its weight from side to side (see Figure 3).

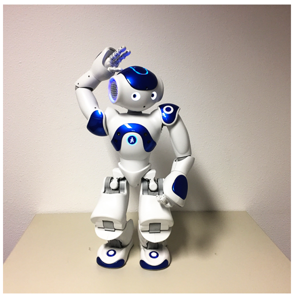


Figure 2: Attempt at showing an iconic gesture for a *rabbit*. The unnatural angle of the arm, positioning of the hand, and movement of the fingers, may lead to confusion and, consequently, adverse effects with respect to learning.

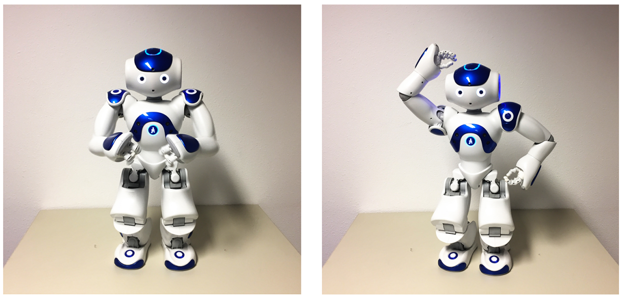


Figure 3: Stills of iconic gestures as depicted by the robot. Left: imitating a *chicken* by simulating the flapping of its wings; right: imitating a *monkey* by simulating the scratching of the head and armpit with the right and left extremities, respectively.

4 RESEARCH QUESTIONS

With the upcoming experiment we intend to answer two research questions. The first question relates to the previous work described in Section 2. We aim to investigate to what extent children will benefit from adaptive language tutoring. We hypothesize an increase in learning gain when children are taught words through an adaptive language tutoring system as compared to a non-adaptive (random) language tutoring system. We anticipate a difference in the exact words that are learned: in the adaptive condition, we expect children to learn those words that were the most challenging during training (having the most incorrect answers) because of the higher repetition rate of these words. In the random condition, the words learned might depend on other factors such as word complexity or attitude towards the animal described by the word.

Our second research question focuses on the effect of gestures on L2 comprehension for children. We hypothesize an increase in learning gain when target words are accompanied by (iconic) gestures during learning, as compared to the absence of gestures. Furthermore, we expect a reduced knowledge decay over time of the words in the gesture condition, similar to the discoveries by Cook et al. in the math problem solving domain with a human tutor [3]. We intend to investigate, using the retention test one week after the experiment, whether these findings carry over to the language learning domain with gestures performed by the robot. It should be noted that participants are not required but also not prohibited from using gestures during the experiment and pre- and post-tests. We are interested in seeing whether children will produce gestures spontaneously following training and, if so, to what extent these gestures will prove to be similar to the ones depicted by the robot.

5 PROPOSED EXPERIMENT

Following the two research questions, our experiment has a 2 (adaptive versus non-adaptive) x 2 (gestures versus no gestures) between-subjects design. We aim to recruit 80 participants, all native Dutch speaking children between the ages of four and five.

Although the proposed experiment is largely a replication of the experiment described in Section 2 and presented in [18], changes to the design had to be made to accommodate the younger participants, as the previous experiment was tailored to adults. Instead of the first interaction between the children and the robot taking place as part of the experiment, the robot will be introduced to the children in a group session the week prior to the experiment to build trust and rapport. We will refer to the robot by a proper name (Robin) and present a background story to stimulate a friendly and open attitude towards the robot [21].

Rather than teaching children the fictional Vimmi words, the target words are the English names of six animals: chicken, monkey, horse, spider, bird, and hippo (used instead of the more difficult hippopotamus). The number of words was reduced to six (from ten in the original experiment, see Schodde et al. [18]) to account for the lower word memory span of children [7], which should be around four words for children of age five. All target words have been selected based on the (varying degrees of) dissimilarity between the words in the L1 (Dutch) and the L2 (English) as well as the feasibility of designing suitable iconic gestures to be performed by the robot to depict the animals. We will conduct a pre-test to verify that participants are familiar with all six target words in their L1 (Dutch) and to test the underlying assumption that participants have no prior knowledge of the target words in the L2 (English). This pre-test will be presented on a different computer screen than the one on which the game is played and without the robot being present, so that there is a clear distinction between this testing environment and the training (game) stage. On the computer screen, the participant will be presented with the pictures of all six animals, one by one. For each picture, the experimenter will ask the participant for the name of the animal in the L1. The computer will then show the pictures of all animals on the screen and name the animals, one after another, in the L2 in random order. Each time the child is prompted with a name in the L2, they are asked to pick the correct image for this animal from the six animals displayed.

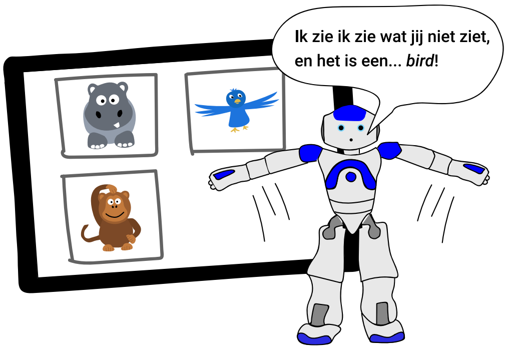


Figure 4: Mock-up of the training phase of the proposed experiment. Three animals appear on the tablet screen, one of which matches the animal picked by the robot. The robot asks the child in their L1 to point out the correct animal based on its name in the L2. In the gesture condition, as shown in this image, the robot performs the associated iconic gesture when mentioning the animal.

The experimental setup uses a Microsoft Surface Pro 4 tablet and the SoftBank Robotics NAO robot. The robot plays a game of "I spy with my little eye", where it picks a certain animal displayed on the tablet screen and names it in the L2, after which the child is expected to tap the corresponding animal picture (see Figure 4). The experimenter inputs the name of the child, so that the robot can personally address the participant, and starts the game. After a brief explanation, the tablet will ask participants to indicate whether they understand the concept of the game. If they indicate that they do not, the experimenter will intervene to provide further explanations. The experiment can be stopped at any time via an experimenter-controlled control panel. Once the actual game commences, the experimenter pretends to be preoccupied so as to avoid participants actively looking for feedback.

In the adaptive learning condition the next target word to train is selected based on the knowledge model (i.e. skill-beliefs) of the participant. After each trial in which the robot exposes the child to one animal, this knowledge model is updated based on the responses of the child. The updated model is then used to select the next target word to be presented. In the random condition, target words are instead randomly presented. In total, there are thirty of these tasks, which means that in the random condition each target word is presented five times throughout the game. In the adaptive condition, the number of times each word occurs depends on how well the participant performs on that specific word, but all words are guaranteed to occur at least once. The previous experiment also consisted of a total of thirty tasks, but as there were ten target words there was less repetition. Reducing the number of words should avoid cognitive overload for the young participants while simultaneously offering more room for the adaptive system to learn the knowledge model of the child and repeat the words that require more attention.

A new addition to the experiment is a condition in which the robot will perform iconic gestures whenever one of the animal names is mentioned in English. These gestures were specifically designed for this experiment, where the robot tries to mimic the appearance or behavior of the animal. The timing of L2 word pronunciation is designed to occur close to the stroke of the gesture. This means that there is a pause in mid-sentence leading up to and after the L2 word, creating additional emphasis on the target. In the condition without gestures, a similar pause is introduced. The robot is set to "breathing mode" in all conditions, which means that it slowly moves its weight from one leg to the other while slightly moving its arms. This prevents the robot from being completely static while, in the gesture condition, reducing the surprise effect of an iconic gesture being triggered.

After thirty prompts to recognize the English animal names, the game finishes. The child is then presented with the post-test, again at the computer screen without the robot. The post-test is identical to the pre-test, except that we no longer test the animal names in the L1. The post-test is also identical across all conditions, so there are no gestures when the L2 words are presented. There are two different images for each animal, one of which will be used for the pre-test and post-test and the other for the game. The images of animals used in the pre-test and post-test feature the same character as those that appear during the game, but in a different pose. The pose in the set of images used during the game is designed to match the gesture that is shown by the robot, to avoid having a mismatch between both sources of visual information for some animal names, and a match for others [23]. For instance, for the word 'bird' the robot will display the act of flying by moving its arms up and down, therefore the bird in the image is also flying. The second set of images could feature the bird facing a different direction, sitting still. By using these two sets of images, we aim to test if children manage to map the English words not only to the specific corresponding image or mental representation of shape, but to the general concept of the animal. One week after the experiment we perform the post-test once again to measure the retention of the six target words.

To assess the iconicity of the gestures, we conducted a perception test with adult participants through an online survey. Participants ($N = 14$) were shown video recordings, one after another, of the six gestures performed by the robot. For each video, participants were asked to answer the question which animal the robot depicted by selecting the corresponding name of the animal in English from a list containing all six target words. The order in which the videos were shown, as well as the order of the items on the list containing the six animal names, was randomized for each participant. Results from the perception test are presented in Table 1. As can be seen from this confusion matrix, with an average accuracy of over 89 percent, participants were, on average, very accurate with respect to their predictions of the depicted gestures. In fact, for three of the six animals (monkey, horse, and bird), not a single mistake was made. With an average accuracy of just over 71 percent, the most ambiguous gestures were those representing the chicken and the hippo. However, it should be noted that participants typically came to realize they had made a mistake, after which they acted accordingly: for example, if a participant was shown the video recording of the chicken prior to that of the monkey and they had incorrectly selected *monkey* as their answer for the recording of the chicken, they would (now correctly) select *monkey* again as their answer when shown the recording of the monkey (we did not allow them to directly correct their past mistake). This implied correction, as well as the high accuracy on average, suggests that we may assume the gestures to be sufficiently iconic, especially as they will ultimately be presented in combination with the verbalization of the name of the associated animal.

Table 1: Confusion Matrix Perception Test

		Perceived					
		Chicken	Monkey	Horse	Spider	Bird	Hippo
Actual	Chicken	10	2	1	0	0	0
	Monkey	0	14	0	0	0	0
	Horse	0	0	14	0	0	0
	Spider	0	0	1	13	0	0
	Bird	0	0	0	0	14	0
	Hippo	1	1	0	2	0	10

Note. Shaded cells indicate true positives.

In our analysis of the experimental results, we intend to measure performance (correct and incorrect answers) during the word training to monitor participants' progress over time in the different conditions. Time on task is measured both in the training "game" and in the post-test. In addition, we will make video recordings of the interaction with the robot for additional analyses (for instance to see if and at what rate children will mimic the robot's gestures). During the post-test we will record how many animals the children managed to correctly identify immediately after training. The retention test will measure decay of the newly attained words after one week.

6 CONCLUSION

The experiment proposed in this paper outlines two valuable topics of discussion for improving the interactions between children and robot, specifically in a tutoring setting. We aim to investigate how the order and frequency of presenting new words in the L2 for the purpose of second language learning can be personalized for each child to optimize learning gain, based on a probabilistic model that traces their knowledge of each word. Second, the experiment evaluates if the positive effect of performing iconic gestures for second language learning by human tutors carries over to the robot.

After running the experiment, future work includes incorporating our findings into the L2TOR project[1]. Adaptive learning will be integrated with the existing lesson plans, improving not only the way the content of each individual lesson is structured but also informing the choice of which words from previous lessons to repeat for better retention. If iconic gestures indeed prove to play a big part in learning and remembering new words, more of these non-verbal behaviors will be developed to accompany a greater number of (target) words and concepts. Furthermore, we will investigate the use of different types of gestures and explore ways of reducing the effort required to implement and orchestrate these gestures for robots. Our progress can be tracked via the project website[1].

7 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). We 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.

REFERENCES

[1] Tony Belpaeme, James Kennedy, Paul Baxter, Paul Vogt, Emiel E.J. Krahmer, Stefan Kopp, Kirsten Bergmann, Paul Leseman, Aylin C. Küntay, Tilbe Göksun, Amit K. Pandey, Rodolphe Gelin, Petra Koudelkova, and Tommy Deblieck. 2015. L2TOR - Second Language Tutoring using Social Robots. In *Proceedings of the International Conference on Social Robotics (ICSR) 2015 WONDER Workshop.*

[2] Kirsten Bergmann and Manuela Macedonia. 2013. *A Virtual Agent as Vocabulary Trainer: Iconic Gestures Help to Improve Learners' Memory Performance.* Springer Berlin Heidelberg, Berlin, Heidelberg, 139–148. DOI: http://dx.doi.org/10.1007/978-3-642-40415-3_12

[3] Susan Wagner Cook, Zachary Mitchell, and Susan Goldin-Meadow. 2008. Gesturing makes learning last. *Cognition* 106, 2 (2008), 1047–1058.

[4] 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. DOI: http://dx.doi.org/10.1007/BF01099821

[5] 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. DOI: http://dx.doi.org/10.1080/1358165042000283101

[6] 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. DOI: http://dx.doi.org/10.1016/j.actpsy.2013.05.013

[7] Frank N. Dempster. 1981. Memory span: Sources of individual and developmental differences. *Psychological Bulletin* 89, 1 (1981), 63–100. DOI: http://dx.doi.org/10.1037/0033-2909.89.1.63

[8] 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. DOI: http://dx.doi.org/10.1007/s10648-015-9334-2

[9] 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. DOI: http://dx.doi.org/10.1080/01690960802365567

[10] 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 Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15).* ACM, New York, NY, USA, 67–74. DOI: http://dx.doi.org/10.1145/2696454.2696457

[11] Sebastian Leitner. 1972. *So lernt man lernen. Der Weg zum Erfolg.* Herder, Freiburg.

[12] Dan Leyzberg, Samuel Spaulding, Mariya Toneva, and Brian Scassellati. 2012. The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci 2012).* Curran Associates, Inc.

[13] 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. DOI: http://dx.doi.org/10.1002/hbm.21084

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

[15] 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. DOI: http://dx.doi.org/10.1016/j.cedpsych.2012.12.001

[16] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. Generation and Evaluation of Communicative Robot Gesture. *International Journal of Social Robotics* 4, 2 (2012), 201–217. DOI: http://dx.doi.org/10.1007/s12369-011-0124-9

[17] Maha Salem, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2011. A Friendly Gesture: Investigating the Effect of Multimodal Robot Behavior in Human-Robot Interaction. In *2011 RO-MAN.* Institute of Electrical and Electronics Engineers (IEEE). DOI: http://dx.doi.org/10.1109/roman.2011.6005285

[18] Thorsten Schodde, Kirsten Bergmann, and Stefan Kopp. 2017. Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making. In *Proceedings of HRI 2017.*

[19] Marion Tellier. 2008. The effect of gestures on second language memorisation by young children. *Gestures in Language Development* 8, 2 (2008), 219–235. DOI: http://dx.doi.org/10.1075/gest.8.2.06tel

[20] 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. DOI: http://dx.doi.org/10.1007/s12369-013-0214-y

[21] Paul Vogt, Mirjam de Haas, Chiara de Jong, Peta Baxter, and Emiel Krahmer. in press. Child-Robot Interactions for Second Language Tutoring to Preschool Children. *Frontiers in Human Neuroscience* (in press).

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

[23] Rolf A. Zwaan, Robert A. Stanfield, and Richard H. Yaxley. 2002. Language Comprehenders Mentally Represent the Shapes of Objects. *Psychological Science* 13, 2 (2002), 168–171. DOI: http://dx.doi.org/10.1111/1467-9280.00430

[1] http://l2tor.eu