Context-Sensitive Natural Language Generation for Robot-Assisted Second Language Tutoring

Bram Willemsen, Jan de Wit, Emiel Krahmer, Mirjam de Haas, Paul Vogt
Tilburg School of Humanities and Digital Sciences, Tilburg University, The Netherlands
{b.willemsen,j.m.s.dewit,e.j.krahmer,mirjam.dehaas,p.a.vogt}@uvt.nl

Abstract

This paper describes the L2TOR intelligent tutoring system (ITS), focusing primarily on its output generation module. The L2TOR ITS is developed for the purpose of investigating the efficacy of robot-assisted second language tutoring in early childhood. We explain the process of generating contextually-relevant utterances, such as task-specific feedback messages, and discuss challenges regarding multimodality and multilingualism for situated natural language generation from a robot tutoring perspective.

1 Introduction

In recent years, an increasing body of work has highlighted the potential of social robots for various educational purposes (Mubin et al., 2013; Belpaeme et al., 2018a). This paper describes research conducted in the context of second language (L2) acquisition in early childhood as part of a project called Second Language Tutoring using Social Robots, or **L2TOR** for short (Belpaeme et al., 2015). The main goal of the L2TOR project is to evaluate the possible benefits of using social robots as (second) language tutors; more specifically, the aim is to provide tentative guidelines to aid the development and deployment of robot-assisted platforms suitable to teach children between the ages of five and six an L2 (Belpaeme et al., 2015, 2018b).

The rationale behind the use of a social robot for the purpose of L2 tutoring is multifold. A noted benefit is the possibility of providing more one-to-one tutoring (Belpaeme et al., 2018a). An advantage of the embodied aspect of a robot tutor is its social and physical presence in the referential world of the learner (Leyzberg et al., 2012). A humanoid robot may capitalise on its anthropomorphic appearance by non-verbally communicating with the learner, such as through the use of gestures, a scaffolding mechanism which has been shown to have positive effects on learning outcomes when used by human tutors (e.g., Hald et al., 2016; Alibali and Nathan, 2007; Tellier, 2008) and may similarly benefit children learning an L2 from a robot tutor (de Wit et al., 2018).

An important aspect in the development of the L2TOR system is the human element; findings from studies of human tutors are leading in the design of the robot's behaviours. The aforementioned use of gestures is an example of nonverbal behaviours to be incorporated into the tutoring interactions. With respect to the verbal behaviours of the robot, the aim is to tailor the lexical output to the situational context of the learner when appropriate. To this end, we turn to natural language generation (NLG). Through context-sensitive NLG, we will be able to provide, among other things, situationally-relevant feedback messages. Adjusting output to fit the situational context is expected to make interactions between child and robot more natural. Situated NLG for human-robot interaction (HRI), however, is a rather complex matter which requires us to address various issues not typically of concern to more conventional applications of NLG. We will discuss in more detail the design choices and challenges encountered with respect to the development of the L2TOR system's multimodal and multilingual output generation module.

2 L2TOR ITS

The L2TOR system is designed to be a state-of-the-art robot-assisted intelligent tutoring sys-

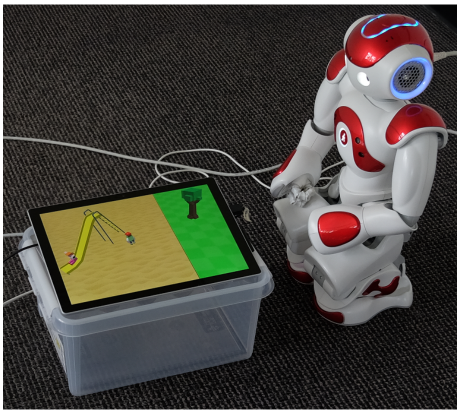

Figure 1: The basic setup of the L2TOR ITS.

tem (ITS) intended to teach young children an L2. The hardware components that constitute the system's learning environment are SoftBank Robotic's NAO humanoid robot and a tablet computer. The basic setup is shown in Figure 1. Combined with promising technologies, such as knowledge tracing (Schodde et al., 2017; de Wit et al., 2018), the motivation behind the system's design is to facilitate the transfer of pedagogical techniques used by human tutors to the robot tutoring domain.

The architecture of the L2TOR ITS is modular in nature; modules in the system are each responsible for dealing with specific parts of the tutoring interaction, including the registration and interpretation of learner inputs, the management of interaction flow and the relaying of relevant information to other parts of the system, and the generation of appropriate behaviours on the basis of knowledge representations derived from learner inputs and situational context. It should be noted that the system relies on the tablet computer to mediate interactions, as automatic speech recognition was considered insufficiently reliable to serve as an input device for child-robot interactions (Kennedy et al., 2017; Belpaeme et al., 2018b).

With the intention of investigating the efficacy of robot-assisted L2 vocabulary training in a longitudinal setting (Belpaeme et al., 2015, 2018b), a series of lessons was developed in conjunction with the L2TOR ITS. The curriculum covered two educational domains, namely the number domain, which involves number words and (pre-)mathematical concepts, and the space domain, which covers basic spatial relations. A total of 34 target words were selected based on a systematic review of educational curricula and standard literacy tests. The lessons were designed to cover, on average, six of the target words per tutoring session. Children were to interact with the robot on seven occasions, i.e., six lessons covering both educational domains followed by a recap session, over the course of roughly three weeks.

3 Generating Output

In the L2TOR ITS, the module responsible for realizing any and all robot output is referred to as the *Output Module*. This output includes both verbal and non-verbal behaviours. Verbal behaviours are realised as synthesised speech through a text-to-speech (TTS) engine. Verbal output is combined with the appropriate non-verbal behaviours such as (co-speech) gestures as well as gaze, all of which is coordinated with accompanying actions on the tablet computer. The Output Module comprises several submodules, each responsible for their own part in the planning and realization of the robot's behaviours. One of these submodules is concerned with the generation of contextually-relevant feedback messages.

The primary purpose of situated NLG for HRI is the contextualisation of output. For a tutoring interaction this means that we would want NLG to be able to take into account the current state of affairs regarding the subject matter as well as the learner's inputs at any point in the interaction to provide them with adequate information, including feedback. In addition, NLG might help make interactions more dynamic by adding variation. Note, however, that certain components, including NLG, in the iteration of the ITS intended to be evaluated in a longitudinal study (Belpaeme et al., 2015, 2018b) are more constrained for reasons of experimental consistency; applications of the system outside of research would ideally increase the level of adaptation and personalization.

3.1 Curriculum

The content of the lessons was designed to provide meaningful context to the target words; in the virtual environment presented on the tablet computer, the children would visit several locations and take part in activities that were related to the language input the child received and which were expected to speak to their imagination. For example, together with the robot, the child would visit the zoo and interact with the animals to learn about numbers and (pre-)mathematical concepts. Activities

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

Figure 2: JSON-formatted data structure containing information regarding current state of the interaction.

```
"monkey": {
    "Dutch": {
        "plural": {
            "article": "de",
            "text": "apen"
        },
        "singular": {
            "article": "de",
            "text": "aap"
        }
    }
}
```

Figure 3: Sample of dictionary containing information on various task-related words and phrases.

then took the form of various tasks. With the tablet in use as the main input device, most of these activities concerned interactions with objects shown on screen (e.g., selecting and moving objects).

The lesson content is stored in so-called *storyboards*. These storyboards are essentially annotated scripts in the form of spreadsheets. They contain line-by-line information regarding expected robot and tablet output at any point in the interaction. Although these storyboards can be amended by non-experts, they are not stored in a machine-readable format. We, therefore, use a custom parser to transform them to a JSON-like format such as shown in Figure 2.

3.2 State Tracking

Even though it is possible to generate contextually-relevant feedback and task descriptions to a certain extent when only the task type and the objects involved are known, this no longer holds when the context requires us to distinguish between several (seemingly) identical objects in order to generate the correct referring expression. For example, this is problematic when a task requires the learner to touch, in the virtual environment on screen, the cage *containing most animals*, but multiple cages are shown. The system will only know that the object associated with task completion is a cage with a specific identifier (ID); this ID is not mapped to any representation that uniquely identifies the object from the others in natural language.

To ensure that the system is aware of which object, in our example *which cage*, was the correct answer, while also being able to generate a description that uniquely identifies it, we implemented a discourse model to keep track of the system's current state — in this case the positions of all virtual objects on the tablet — over the course of the interaction. To make sure that these object descriptions are generalizable to different languages and various situations, the model stores data structures, such as shown in Figure 2, instead of full utterances. The components of this data structure (cage, containing, most, animals) can then be translated using a dictionary, such as shown in Figure 3, before being inserted into the correct syntactic template. The conversion between object IDs and their descriptions is currently performed offline, i.e., prior to the interaction rather than during, when parsing the storyboards. During the interaction, the discourse model is supplemented by functionalities from Underworlds (Lemaignan et al., 2018), a spatial and temporal modelling framework, which tracks, in real time, whether certain tasks have been correctly carried out in the virtual environment.

3.3 NLG

As a result of the task-driven and scripted nature of the tutoring interactions, NLG serves a niche purpose within the ITS. Although progress has been made with respect to end-to-end NLG systems (Gatt and Krahmer, 2018), given the focused domain of application, namely situated NLG for robot-assisted L2 acquisition, we have instead opted for a template-based approach (van Deemter et al., 2003; Gatt and Krahmer, 2018) as this allows us to exert the necessary control over the output, both verbal and non-verbal, to ensure its quality. Similarly to other data-to-text systems (Gatt and Krahmer, 2018), we use hand-crafted syntactic templates and fill gaps with task-specific information. This information is derived from data structures such as shown in Figure 2 and Figure 3.

Part of the interaction for which NLG is re-

“Nee, dat klopt niet helemaal.	[No, that’s not quite right.]
Je moet **the monkey** in de kooi *aanraken.*	[You need to touch **the monkey** in the cage.]
Probeer het nog maar een keer.”	[Try again.]

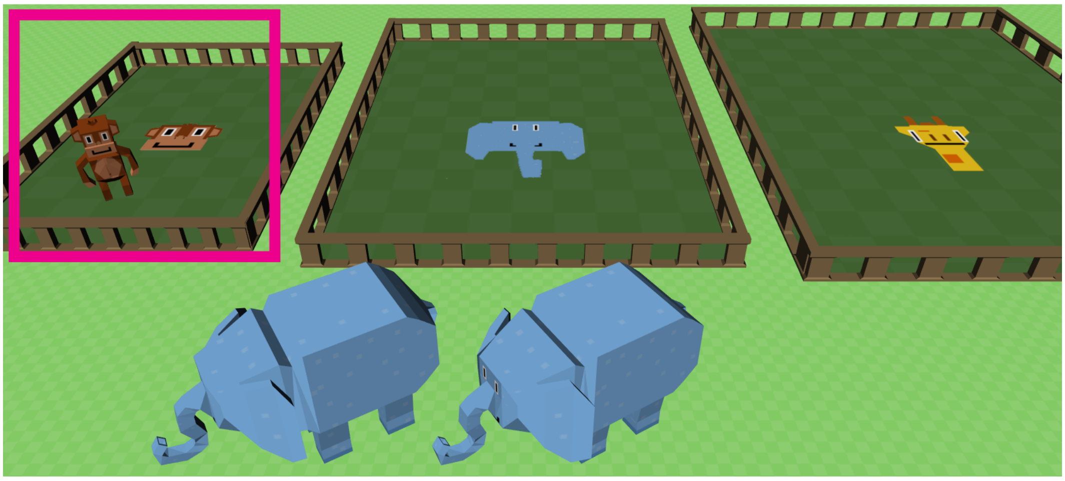

Figure 4: Example of an object selection task presented on the tablet computer. For this hypothetical scenario we assume the L1 to be Dutch and the L2 to be English. The learner was asked to touch the monkey in the cage — here shown in the (pink) box —, but has instead touched one of the elephants shown in the foreground. The robot will now provide the feedback message as shown above the image (in italics). Note that neither the (pink) box nor the text are visible to the learner.

quired is the contextualisation of feedback. Rather than telling the learner their execution of a task was either wrong or right, we want to be able to comment on the exact nature of their actions in relation to what was required of them for task completion. The information required to make feedback messages contextually relevant varies per task, as does the way in which this information is organised. For this reason, different tasks require the use of different syntactic templates for the provision of adequate feedback.

To illustrate the process of constructing a contextually-relevant feedback message, Figure 4 provides an example of an incorrectly-executed object selection task in an interaction in which the L1 is Dutch. At this point in the interaction, there are several animals shown on screen, one of which is shown inside an enclosure (referred to as cage), namely the monkey; two elephants, however, have managed to escape. The learner is asked to touch the monkey residing inside the cage, but does not manage to do so. In order to provide feedback to the learner, we use the template as shown in Table 1. The template contains a preposition (**$prep**) explaining the relationship between two objects, here labeled as **$trg** (target) and **$obj** (object). In our example, the target is the noun phrase *the monkey* and the object is the noun phrase *the cage* (*de kooi* in Dutch). In order to retrieve the correct form, we consult a dictionary with information regarding the objects in question, such as shown in Figure 3. If the target in our example had been addressed in the L1, we would have retrieved the Dutch singular version of the noun phrase, i.e., the determiner *de* [the] and the noun *aap* [monkey]. To complete the feedback message, the syntactic template is preceded by a feedback phrase indicating more explicitly that an incorrect answer was provided, and followed by a prompt telling the learner to attempt the task once more. Although the prompt is hard-coded, the feedback phrase concerns a random selection, without immediate repetition, from a set of canned expressions as a way of introducing some more variation to the message.

In addition, in the event that user input is not registered for an extended period of time, we attempt to re-engage the learner through a contextually-relevant prompt. This prompt is constructed in a similar manner as the feedback message, i.e., by means of slot-filling a task-relevant syntactic template, to remind the user of the current task.

3.4 Non-Verbal Behaviour

Human tutors often use gestures as a scaffolding mechanism (e.g., Alibali and Nathan, 2007).

`(A)`	Nee, dat klopt niet helemaal.	[No, that's not quite right.]
`(B)`	Je moet **`$trg $prep $obj`** aanraken.	[You need to touch **`$trg $prep $obj`**.]
`(C)`	Probeer het nog maar een keer.	[Try again.]

Table 1: Example of a feedback message for an object selection task. The message consists of three parts: `(A)` a (negative) feedback phrase, `(B)` the syntactic template, and `(C)` a prompt.

Thanks to the NAO's humanoid appearance, we can incorporate gestures into tutoring interactions in a similar manner. For gestures that coincide with speech, i.e., co-speech gestures, the proper alignment of speech and gesture is crucial. This behavioural management is a built-in functionality of the NAOqi API. The ALAnimatedSpeech module[1] processes text annotated with specific commands in order to tell the robot at which point in an utterance a behaviour, such as an iconic gesture, is to be executed. To improve the timing of the execution, we inserted timed pauses to synchronise the stroke of the gesture with the target word. Despite increased synchronisation, the added pauses do slow down the interaction.

In addition to iconic gestures, we make use of deictic gestures to guide the learner's attention. The combination of gaze and pointing gestures helps establish joint attention, while gaze may also help build rapport between child and robot (Admoni and Scassellati, 2017). All non-verbal behaviours are triggered from the annotated utterance, of which an example is shown in Table 2.

3.5 Speech Synthesis

In contrast with typical NLG applications, the surface realization of NLG for HRI is not a human-readable text, but instead a rendition of an utterance as synthesised speech. Depending on the language of choice, the TTS engine of the NAO robot is by default either powered by Nuance or Acapela. These TTS engines are both capable of producing a speech signal from a text string.

In the context of language acquisition, the quality of the synthesised speech may be of importance, as (young) learners have been shown to attend to non-verbal cues present in the speech signal when presented with a novel language (e.g., Dominey and Dodane, 2004). Although the effects of speech synthesis quality on learners' perceptions have previously been studied for computer-assisted language learning (e.g., Bione et al., 2016; Handley, 2009; Kang et al., 2008), whether poor quality speech synthesis impedes the efficacy of language acquisition has not been unequivocally established.

Although both the Nuance and Acapela TTS engines allow for modification of the speech signal to a certain extent by means of parameter tuning (e.g., pitch, volume, speaking rate), control over the quality of the synthesised speech is limited. The multilingual nature of the interaction causes additional difficulties, as code-switching in the current iteration of the ITS requires us to switch TTS engine frequently, often within the same utterance. As a result of the engines only receiving segments of the utterance rather than the utterance as a whole, the quality of the speech signal is negatively affected as words and phrases, in particular near segmentation boundaries, are mispronounced to varying degrees. It should be noted that the switch of engine also results in a change of voice, as different languages have been dictated by different speakers.

Despite certain difficulties being inherent to the technologies themselves, we have managed to address some of the TTS problems we have encountered. For example, in order to correct some of the pronunciation errors, we have relied on phonetic transcriptions of problematic words and phrases. Take, for instance, the word *tablet*. When the L1 is Dutch, the TTS will pronounce the word as the Dutch word for *pill*, rather than the intended pronunciation referring to a tablet computer. However, when we use the following phonetic representation of the word: `t E: b l @ t`, the synthesised speech will more closely resemble the expected pronunciation. Furthermore, to avoid any chance of poorly synthesised speech being a learner's first exposure to a target word in the L2, we instead make use of audio recordings of a native speaker, played back via the tablet's speakers.

4 Conclusion

In this paper, we have described the L2TOR ITS, focussing primarily on the system's multimodal

[1] http://doc.aldebaran.com/2-1/naoqi/audio/alanimatedspeech.html

Kijk **John** `^start(pointing/tablet)` `$toggle_facetracking=False` `^start(gaze/tablet)` ,
de dieren spelen een spelletje met ons! `$toggle_facetracking=True`

[Look **John** `^start(pointing/tablet)` `$toggle_facetracking=False` `^start(gaze/tablet)` ,
the animals are playing a game with us! `$toggle_facetracking=True`]

Table 2: Example of an annotated utterance returned by the Output Module. Here, **John** is the child's given name. **^start(pointing/tablet)** indicates that the robot will direct the attention of the child to the tablet by using a pointing gesture. As can be seen from **$toggle_facetracking=False**, face tracking is then disabled, after which the robot will direct its own gaze towards the tablet, **^start(gaze/tablet)**, in an attempt to establish joint attention. At the end of the utterance, face tracking is once again enabled.

and multilingual output generation module. We have discussed challenges with respect to situated NLG for the purpose of robot-assisted language tutoring, including natural-sounding TTS, multimodality and multilingualism, coordinating robot actions and tablet output, and how and to what extent these were addressed within the context of the project.

Acknowledgements

This work is funded by Horizon 2020, the EU Framework Programme for Research and Innovation, Grant Agreement: 688014, and the Tilburg School of Humanities and Digital Sciences at Tilburg University, The Netherlands.

References

Henny Admoni and Brian Scassellati. 2017. Social Eye Gaze in Human-Robot Interaction: A Review. *Journal of Human-Robot Interaction*, 6(1):25–63.

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):349–366.

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 First International Workshop on Educational Robotics at ICSR 2015*.

Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018a. Social robots for education: A review. *Science Robotics*, 3(21):eaat5954.

Tony Belpaeme, Paul Vogt, Rianne van den Berghe, Kirsten Bergmann, Tilbe Göksun, Mirjam de Haas, Junko Kanero, James Kennedy, Aylin C. Küntay, Ora Oudgenoeg-Paz, Fotios Papadopoulos, Thorsten Schodde, Josje Verhagen, Christopher D. Wallbridge, Bram Willemsen, Jan de Wit, Vasfiye Geçkin, Laura Hoffmann, Stefan Kopp, Emiel Krahmer, Ezgi Mamus, Jean Marc Montanier, Cansu Oranç, and Amit Kumar Pandey. 2018b. Guidelines for Designing Social Robots as Second Language Tutors. *International Journal of Social Robotics*, 10(3):325–341.

Tiago Bione, Jennica Grimshaw, and Walcir Cardoso. 2016. An evaluation of text-to-speech synthesizers in the foreign language classroom: learners' perceptions. In *CALL communities and culture short papers from EUROCALL 2016*, pages 50–54.

Kees van Deemter, Mariët Theune, and Emiel Krahmer. 2003. Real vs. template-based natural language generation: a false opposition? *Computational Linguistics*, 31:15–24.

Peter F. Dominey and Christelle Dodane. 2004. Indeterminacy in language acquisition: the role of child directed speech and joint attention. *Journal of Neurolinguistics*, 17(2-3):121–145.

Albert Gatt and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

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):495–522.

Zöe Handley. 2009. Is text-to-speech synthesis ready for use in computer-assisted language learning? *Speech Communication*, 51(10):906–919.

Min Kang, Harumi Kashiwagi, Jutta Treviranus, and Makoto Kaburagi. 2008. Synthetic speech in foreign language learning: an evaluation by learners. *International Journal of Speech Technology*, 11(2):97–106.

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. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 82–90.

Séverin Lemaignan, Yoan Sallami, Christopher Wallbridge, Aurélie Clodic, Tony Belpaeme, and Rachid Alami. 2018. underworlds: Cascading Situation Assessment for Robots. In *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Daniel 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*, pages 1882–1887.

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 Learning*, 1(1):1–7.

Thorsten Schodde, Kirsten Bergmann, and Stefan Kopp. 2017. 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*, pages 128–136.

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

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 *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 50–58.