Social robots as second-language tutors for young children

Challenges and opportunities

Rianne van den Berghe

Cover	Leonie Vlaar
Photo	Broosdoc Films
Printing	Ridderprint

ISBN	978-94-6375-412-5

Social robots as second-language tutors for young children

Challenges and opportunities

Sociale robots als tweede-taaltutors voor jonge kinderen

Uitdagingen en kansen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op
vrijdag 14 juni 2019 des middags te 2.30 uur

door

Maria Anna Josina van den Berghe

geboren op 20 maart 1993
te Schagen

Promotor:

Prof. dr. P.P.M. Leseman

Copromotoren:

Dr. J. Verhagen

Dr. O. Oudgenoeg-Paz

This dissertation was accomplished with financial support from the European Union Horizon2020 program (grant number 688014).

Contents

1

General introduction

Technology plays an increasingly important role in our everyday lives. Many people own a smartphone that they use to communicate with people all over the world, to control their household appliances from a distance, or to stay updated on the world's news. Newly developed technologies also have found their way into the field of education. For example, interactive white boards, chat programs, and instructive virtual games are increasingly used in education for both children and adults (Golonka, Bowles, Frank, Richardson, & Freynik, 2014; Takacs, Swart, & Bus, 2015; Young et al., 2012). This dissertation focuses on the use of technology in education, in particular second-language (L2) education for children.

An important and timely question in education is how technology can contribute to students' learning of an L2. Technologies have some advantages that are not present in traditional classrooms (Golonka et al., 2014). Imagine, for example, a child who is a starting L2 learner and who is still struggling with the new language and needs more feedback and personalized instruction. The teacher cannot always meet the needs of the child, as the rest of the class is waiting for the teacher's attention too. A computerized adaptive language teaching program, instead, may enable the child to engage in additional well-tailored L2 learning activities without intensive involvement of the teacher. Consider another example. Technology may provide native language level input to children in virtually all languages in the world. This may be especially helpful for migrant children who learn a different language at home than the language spoken at school. There is increasing evidence that building on children's first language (L1) in education supports balanced bilingual development (Blom, 2019). However, building on children's L1 requires teachers who know these languages and can provide native-level input in these languages. Clearly, teachers may speak some of children's L1s at an adequate level of proficiency, but this is rather exceptional and with the increasing linguistic diversity in classrooms not a real option. Chat programs, interactive animated story books, educational games and other technology-supported educational tools can be a solution.

Optimizing L2 education is important for several reasons. The value current society attaches to learning several languages is increasing, due to globalization. Nowadays schools, and even preschools and daycare centers, are offering additional languages, mostly English in the Dutch context, but also other languages to children already from an early age (Nikolov & Djigunović, 2006). The European Union has declared that every European citizen should be taught practical skills in at least two languages other than the mother tongue starting already at a young age (BEC, 2002). Yet, implementation of this policy still faces many difficulties. Migration within Europe and from outside Europe has increased the past decades, resulting in increasing numbers of children with highly varied L1s in classrooms, a phenomenon referred to as 'linguistic superdiversity' (Vertovec, 2007). All these children have to master the school language as the second or sometimes third or fourth language as quickly and efficiently as possible to prevent education gaps. It is important for these children to be supported in learning both the national language as well as their native language to promote balanced bilingual development, subject learning, self-confidence and wellbeing (Blom, 2019; Cummins 2008; Creese & Blackledge, 2010; Hornberger, 2005). All these developments together pose serious challenges for education and technology may aid in tackling at least some of them.

Social Robots in Education

Recently, a new form of technology has been introduced to education: social robots. Social robots are robots that are specifically designed to interact with people. They can perform tasks while being controlled by a person in real-time (i.e., semi-autonomous robots) or through predefined scripts which allow the robots to engage in interactions without a controller (i.e., autonomous robots). In interactions, social robots follow behavioral norms that are typical for human interaction and they can make use of behaviors that are inherent to human communication, such as coordinating eye gaze, pointing, and other types of gestures (Bartneck & Forlizzi, 2004). Robots are already widely used in industry (e.g., robotic arms in factories) and even at people's homes (e.g., robotic vacuum cleaners). Such robots, however, are designed to perform a specific task without engaging in complex social interactions with people. Unlike these robots, social robots have a humanoid appearance and can dispose of speech recognition and production devices, sensors, limbs, and motor abilities that enable, at least in theory, a more natural interaction with humans than is possible with other forms of technology.

While the ability to use eye gaze, pointing, or other types of gestures also holds to some extent for virtual agents presented in 3D on a computer screen or tablet, social robots crucially differ from virtual agents in that they have a physical body. Social robots are present in the real world, which brings, at least in theory, additional advantages. Being present in the world enables robots to interact with the real-life and real-time physical environment that is shared with the human interaction partners. Social robots can move through the environment, manipulate objects, and, in educational interactions with human learners, establish joint attention, create a common ground, collaborate and physically get in touch with the human learner, which virtual agents cannot do. These are the promises.

According to recent insights, embedding communicative interactions in the shared physical environment is especially important for young children's language development and may also be important for L2 learners (Barsalou, 2008; Hockema & Smith, 2009; Iverson, 2010; Oudgenoeg-Paz, Volman, & Leseman, 2015; Wellsby & Pexman, 2014). Several studies have pointed to the importance of gestures to link the physical experiences of the conversational partners when perceiving or acting upon objects and events in the world to the language-to-be-learned (Macedonia, Müller, & Friederici, 2011; Rowe, Silverman, & Mullan, 2013). Social robots can attend to the same objects as children, act in the shared world in coordination with them, and use pointing and iconic gestures to make meaning clear, while simultaneously uttering speech in the relevant language to map language onto its referents. Thus, robots can potentially provide the type of interaction that is known to promote children's language learning. These are the possibilities.

Robots can display various behaviors that have shown to benefit L2 learning specifically, at least when performed by humans. Previous research on L2 vocabulary learning, for example, has shown that learners benefit from rich explicit instruction supported by pointing, gesturing, the use of physical objects, and acting-out scenarios (Collins, 2010; Jalongo & Sobolak, 2011). Both deictic and iconic gestures aid in children's L2 learning (Rowe et al., 2013; Tellier, 2008). Also, the use of students' L1 to explain the meaning of L2 words has been found to contribute to L2 learning (Jiang, 2005; Carlo et al., 2004). Social robots can, in principle, realize all this, as they can point, gesture, act, and switch between multiple languages. The possibility to speak any language, the multimodality, and the physical presence are the main reasons why social robots could be more effective in L2 education than other forms of technology. These are the expectations.

Even though the potential advantages of robots are clear and seem to fit in well with recent insights on the nature of language acquisition and L2 learning, still very little is known about the actual effectivity of social robots in interventions to support language learning (for recent reviews, see Chapter 2 of this dissertation; Kanero et al., 2018). Much is still unclear regarding both the optimal design of human-robot interactions for L2 learning and the effectivity of L2 education involving social robots. This dissertation aims to shed light on these issues by investigating key principles in designing robot-assisted language learning (RALL) interactions and by assessing the added value of robots for L2 education.

Children's L2 Word Learning with Social Robots

Studies on L1 vocabulary training programs show that such interventions can be highly effective, according to a recent systematic review, with an average effect of nearly one standard deviation on measures of the trained vocabulary (Marulis & Neuman, 2010). According to this review, a combination of explicit and implicit instruction is the most effective pedagogical approach to teach new vocabulary, as learners are first engaged in explicit instruction of the new words and can then subsequently apply them in a broader context of meaningful practice. Another characteristic of effective vocabulary training programs is even more important for their success, that is, the way in which the person who delivers the instruction is sensitive to the learners' understanding and support needs, adapts to the learners, signals interest, invites production, and responds contingently to learners' utterances. Effective vocabulary training programs, therefore, are without exception characterized by extensive training of the teachers or other intervention agents to ensure optimal implementation.

The latter finding is of particular interest in the context of RALL, as it raises a fundamental question. If even experienced teachers need intensive training to be sensitive, adaptive, and responsive, and to behave contingently in vocabulary training interventions with children, can a robot, then, ever succeed in teaching new vocabulary to children? Few studies to date have actually investigated L2 word learning with social robots and found mixed results (de Wit et al., 2018; Gordon et al., 2016; Kanda, Hirano, Eaton, & Ishiguro, 2004; Kory Westlund et al., 2015; Mazzoni & Benvenuti, 2015; Tanaka & Matsuzoe, 2012), which may stem from two unresolved issues.

The first issue concerns the number of sessions in robot-assisted vocabulary training programs. Two studies in which children played games with a robot in a single session showed that children learned, respectively, three out of six and two to three out of four target words (de Wit et al., 2018; Tanaka & Matsuzoe, 2012), which as such seems reasonable (but note that the sessions pertained to a small number of target words only). In contrast, two studies in which children played games with a robot either as much as they wanted during a period of two weeks (Kanda et al., 2004) or during eight sessions (Gordon et al., 2016), found that children performed at best just above chance level on post-test comprehension tests. Based on the research into vocabulary training programs with a human teacher (Marulis & Neuman, 2010), the opposite pattern would have been expected, as another characteristic of effective vocabulary trainings is that they involve multiple sessions spread over a period of several weeks to several months, present a relatively large number of target words that are coherently connected, build-up over sessions, and cover well-defined domains of conceptual knowledge. Why is this apparently different in RALL, as studied so far?

A possible explanation is the novelty of the robot. People are often not used to playing or working with robots, and may be excited and focus more on the robot (and subsequently learn more) than when they would have had more experience with playing or working with robots (see Leite, Martinho, & Paiva, 2013, for an overview of long-term interactions with robots). To put it differently, the enhanced motivation and attention due to the novelty of the robot may have temporarily increased the learning effects in children (the finding of single session studies), but the decline in motivation in subsequent sessions and the much smaller learning effects, if any, as a consequence, may be more representative for the true effectivity of RALL, which seems limited (the finding in multiple session studies). Therefore, to truly test the effectivity of robots beyond the effect of the initial novelty, studies involving multiple interaction sessions are essential.

There is another reason for the need to implement a multiple sessions intervention. Single sessions train only a few words. Yet, for RALL to have real impact and to serve the needs of L2 learning as they exist in the reality of education, the ambitions should be set higher and RALL should be extended to comprehensive vocabulary training programs that can effectively train a large number of words, much like the successful traditional vocabulary training programs reviewed by Marulis and Neuman (2010).

Second, well-known in the educational sciences are so called treatment-aptitude effects in educational interventions. Taking treatment-aptitude effects into account is useful to identify for which learners the intervention is effective and for which it is not, or less so, and may lead to adaptation of the intervention. Treatment-aptitude studies are rare in RALL research. Yet, given the complex multimodal and dynamic nature of language learning involving robots, in combination with differences between children in abilities and skills relevant for this task, it is likely that there are treatment-aptitude effects in RALL as well. Some children may benefit from a robot when learning L2 words, while others may not, or less so. Currently, it is not clear whether and how individual child characteristics moderate the effectivity of RALL.

Finally, relatively little is known about the effectivity of robots compared to non-RALL interactions, such as learning from human speakers or through other types of technology. One study compared children's L2 word learning when children were taught by an adult teacher, a tablet, or a social robot. Although children preferred being taught by the robot, no differences in learning gains were found between the conditions (Kory Westlund et al., 2015). Only a few RALL studies have compared children's robot-assisted learning to children's learning using a tablet only, focusing on reading skills. These studies found a benefit for the robot-assisted condition (Gordon, Breazeal, & Engel, 2015; Han, Jo, Jones, & Jo, 2008; Hyun, Kim, Jang, & Park, 2008). With respect to word learning, however, it is not yet clear whether robots can have an added value for education.

The L2TOR Project

The current dissertation was carried out within the L2TOR project, an international research and development project funded by the European Union within the Horizon2020 program[1]. The main aim of the project was to develop and test a social robot that can aid in young children's L2 word learning. A complete robot system was developed (see Figure 1), applied and further refined throughout the project, and evaluated in a large-scale randomized controlled trial at the end of the project. Children played language games together with a NAO robot[2]. A tablet was used as a mediating device between the robot and the child, while the educational content was displayed on the tablet. The tablet was used as a mediating device because, given the current state of technology, the robot's speech recognition was not capable of

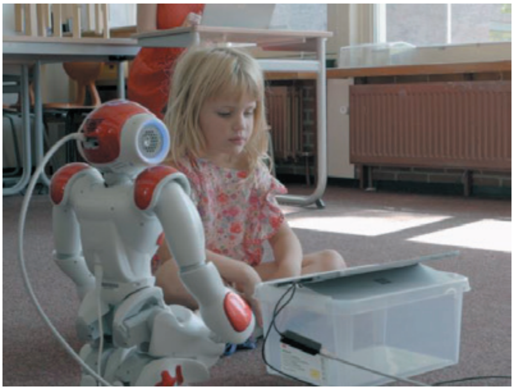


Figure 1. A child playing L2 vocabulary games with the L2TOR setup.

[1] www.l2tor.eu

[2] Developed by Softbank Robotics, see https://www.softbankrobotics.com/emea/en/robots/nao.

recognizing and accurately processing speech of young children (Kennedy et al., 2017), nor to manipulate physical objects (Wallbridge, Lemaignan, & Belpaeme, 2017). We will return to this issue in the General Discussion.

Various lines of research were combined in the L2TOR project. For example, the benefits of adaptivity and different types of gestures (such as iconic or deictic) on learning were studied to design a more effective RALL interaction situation (de Wit et al., 2018; Schodde, Bergmann, & Kopp, 2017). Children's engagement during RALL sessions and the effects of various types of feedback were studied to gain insight into what would be the most effective feedback in RALL interactions. Also, efforts were made to enable the robot to understand and produce spatial language to allow for natural interactions that would benefit learners. Part of these efforts are reflected in this dissertation. Complete information can be found in various reports available through the project's website and in several scientific publications.

This Dissertation

This dissertation reports a series of studies that reflect the role of Utrecht University as partner in the L2TOR project. We developed together with colleagues at Koç University, Turkey, a lesson series for L2 word learning based on current knowledge about the effective characteristics of vocabulary training programs. The lesson series was implemented in the L2TOR system. Target words were chosen such that they were part of early academic language, more specifically early mathematical and spatial language, in line with the recommendation to select target words at the so called tier 2 level (Beck & McKeown, 1985) within coherent domains of conceptual knowledge (Marulis & Neuman, 2010). Early academic language is highly important for later academic success (Leseman, Henrichs, Blom & Verhagen, 2019; Esser, 2006; Hoff, 2013). Although especially children with a different L1 than the school language need support in acquiring academic language, a focus on tier 2 academic language vocabulary seems an efficient strategy in general, as knowledge of this type of words can help children to get access to and disambiguate ambient discourse, and thereby stimulate spontaneous implicit learning of other words (Beck & McKeown, 1985). Note that in this dissertation only studies in which Dutch children learned English as an L2 through the L2TOR system are presented. A separate study using the L2TOR system was conducted among bilingual Turkish-Dutch immigrant children learning Dutch as L2 (Leeuwestein et al., in prep.)

The studies included in this dissertation address several aspects of the effective design of human-robot interaction for preschoolers and were conducted in the course of the L2TOR project to inform the development of the final L2TOR system. The main aims of this dissertation were to (1) synthesize earlier findings on the use and effectiveness of social robots for language learning, and identify in particular current issues and avenues for future research; (2) investigate the use of tablets in RALL to inform the design of RALL interactions; (3) investigate the added benefits of robots for L2 word learning; and (4) investigate whether a robot's benefits depend on differences between children in language and attention skills and their perception of the robot.

Chapter 2 presents a review on the use of social robots for language learning. Thirty-three studies on RALL, targeting different languages, age groups, and aspects of language, and using different robots and methodologies, were selected and discussed. Besides insights into learning gains obtained in RALL situations, these studies raise general questions regarding students' motivation and robots' social behavior in such learning situations.

Chapter 3 addresses the use of tablets with 3D representations of objects versus the use of real physical objects. Tablets are often used as a mediating device in RALL, as most social robots are not yet capable of manipulating physical objects, but it is not clear how this affects learning. From an embodied cognition perspective (Barsalou, 2008; Hockema & Smith, 2009; Iverson, 2010; Wellsby & Pexman, 2014), one would expect that children learn more when interacting with physical objects rather than with virtual objects on a tablet screen. Thus, RALL interactions could be more effective when using physical objects than working with virtual objects displayed on tablets. To settle this issue, we compared the effects of physical versus virtual objects on a tablet screen on children's L2 word learning.

Chapter 4 and 5 present two studies assessing the added value of robots for L2 learning. In the study of Chapter 4, we compared settings in which children played language games on a tablet without a peer, with a child peer, or with a robot. This comparison was not only useful in a theoretical perspective, but also in view of education practice and policy, as a solid proof of clear added value of robot-assisted interactions over interactions without robots is needed before large-scale implementation of robots in schools can be recommended.

Chapter 5 presents our main study within the L2TOR project: a large-scale randomized controlled trial into the effectiveness of robots as L2 tutors involving an English as L2 vocabulary training program, consisting of multiple sessions and conducted among 193 children in nine kindergartens in the Netherlands. Studies of this scale are extremely rare in RALL research and this study allowed us to investigate the effectiveness of the robot in a realistic context – the context in which the robot should be implemented ultimately. This study also addresses another key question in RALL research, namely the comparison of a combined robot-tablet setup to the use of a tablet only in an L2 vocabulary training consisting of multiple sessions.

In addition, Chapter 5 and 6 each discuss differences between children that may moderate, as in treatment-aptitude interactions, the degree to which they learn from the robot. In the main randomized controlled trial, we not only investigated the added value of the robot but also whether differences between children in skills relevant for language learning affected the degree to which they learned from the robot. Specifically, we investigated whether differences between children in vocabulary knowledge in their L1, phonological memory, and selective attention (reported in Chapter 5) and the degree to which children attributed human-like characteristics to the robot (reported in Chapter 6) moderated the effectivity of the robot.

The dissertation concludes with a general discussion in Chapter 7, in which we reflect on the findings of the various studies and on the future of robots in (language) education.

2

Social robots for language learning: A review

Rianne van den Berghe, Josje Verhagen, Ora Oudgenoeg-Paz, Sanne van der Ven, & Paul Leseman. (2019). Social robots for language learning: A review. *Review of Educational Research, 89*(2), 259-295.

Author contributions: R.B., J.V., O.O., S.V., and P.L. conceptualized the research; R.B. conducted the literature search and wrote the paper; J.V., O.O., S.V., and P.L. supervised and critically reviewed the paper.

Abstract

In recent years, robots have increasingly been implemented as tutors in both first- and second-language education. The field of robot-assisted language learning (RALL) is developing rapidly. Studies have been published targeting different languages, age groups, and aspects of language, and using different robots and methodologies. The present review presents an overview of the results obtained so far in RALL research and discusses the current possibilities and limitations of using social robots for first- and second-language learning. Thirty-three studies in which vocabulary, reading skills, speaking skills, grammar, and sign language were taught are discussed. Besides insights into learning gains attained in RALL situations, these studies raise more general issues regarding students' motivation and robots' social behavior in learning situations. This review concludes with directions for future research on the use of social robots in language education.

Keywords: robot-assisted language learning, human-robot interaction, first- and second-language learning, motivation, novelty effect

Technologies such as computers, tablets, and smartphones offer a wide array of possibilities for first- and second-language learning. These forms of technology, in particular interactive white boards, automatic speech recognition programs, instructive virtual games, chat programs, tablets, and animated books, are increasingly being integrated into language education for both children and adults (Golonka, Bowles, Frank, Richardson, & Freynik, 2014; Takacs, Swart, & Bus, 2015; Young et al., 2012). These technologies allow for forms of language learning that are not always present in traditional classrooms, such as one-to-one and tailored instruction, access to native language input, direct feedback, and the possibility to practice with a virtual agent, which may be less intimidating than practicing with a peer or a classmate (Golonka et al., 2014).

One of the newest forms of technology used in education—and the focus of the present review—are social robots. Social robots are robots that are specifically designed to interact and communicate with people, either semi-autonomously or autonomously (i.e., with or without a person controlling the robot in real-time), following behavioral norms that are typical for human interaction (Bartneck & Forlizzi, 2004). These robots are different from, for example, robotic arms in factories, which are often designed to perform a specific task and generally do not interact with people. They also differ from virtual agents or computer-based intelligent tutoring systems, as social robots always have a physical body of some sort and are, therefore, present in the real world, rather than being only virtually present via a screen. The field of robotics has developed rapidly over the last decade, leading to the availability of robots that can be employed for educational purposes. In recent experiments, robots have been used as tutors, for example in teaching prime numbers (Kennedy, Baxter, Senft, & Belpaeme, 2015), puzzle-solving skills (Leyzberg, Spaulding, Toneva, & Scassellati, 2012), and, even more recently, in teaching language (e.g., Alemi, Meghdari, &

Ghazisaedy, 2014; Kennedy, Baxter, Senft, & Belpaeme, 2016). The main aims of this review are to present the current state of knowledge about robot-assisted language learning (RALL), discuss advantages and disadvantages of RALL, and identify potential areas for future research on this topic.

Robots are presumed to have at least two advantages over most other forms of technology. First, they allow the learner to interact with the real-life physical environment, which is thought to be important for language development (Barsalou, 2008; Hockema & Smith, 2009; Iverson, 2010; Wellsby & Pexman, 2014). Specifically, both the manipulation of real-life objects (Kersten & Smith, 2002) and the use of whole-body movement and gestures (Mavilidi, Okely, Chandler, Cliff, & Paas, 2015; Rowe & Goldin-Meadow, 2009; Toumpaniari, Loyens, Mavilidi, & Paas, 2015) have been found to help children's vocabulary learning. Because of the possibility of acting on the physical environment, robots offer possibilities not provided by traditional computer-assisted lessons, such as manipulating objects and using gestures to support language teaching (e.g., Alemi et al., 2014).

The second advantage is that robots allow for more natural interaction than other forms of technology because of their appearance, which is often humanoid or in the shape of an animal. Many robots can use non-verbal cues such as eye gaze, pointing, and other types of gestures. While this also holds for animated characters on a screen, robots are generally perceived as more helpful, credible, informative, and enjoyable to interact with than animated characters (Kidd & Breazeal, 2004; Wainer, Feil-Seifer, Shell, & Matari, 2007). Furthermore, robots are more likely to be perceived as a typical teacher, peer, or friend rather than as a machine: Both children and adults have a tendency to anthropomorphize robots, that is, to ascribe human-like characteristics and behaviors to robots (Bartneck, Kulić, Croft,& Zoghbi, 2009; Beran, Ramirez-Serrano, Kuzyk, Fior, & Nugent, 2011; Duffy, 2003). Therefore, robots can be programmed to take up a specific role, for example the role of a teacher or a friend, depending on whether the aim of the learning tasks is to instruct or correct students on a task, or to have them practice newly learned information with peers.

Even though it is clear which advantages robots potentially have, there are a number of issues that need to be addressed in order for robots to be effective language tutors (see also Kanero et al., 2018, for a review on early language learning). The present review presents the current state of RALL research, with a special focus on affective aspects such as students' motivation and their responses to robots' social behavior. The overall goal of our review is to gain insight into the potential of robots as first- and second-language tutors and to identify areas for further research. Studies on preschool children, school-aged children, and adults will be reviewed. Throughout our review, studies will be described in relative detail to allow a thorough evaluation of the studies conducted and the possibilities robots offer for supporting language learners.

Our review is organized as follows. First, we describe our search criteria and the studies that were selected for review. Second, we present studies focusing on the effects of RALL on participants' language-learning outcomes. In these studies, word learning has been investigated more extensively than other aspects of language and will be discussed first, followed by a discussion of RALL studies on grammar learning, reading skills, speaking skills, and sign language. Third, we describe studies focusing on the role of affective aspects of RALL, addressing how robots may affect learners' motivation, the role of the robot's novelty, and the effect of robots' social behavior on learning. Finally, we discuss the meaning of these findings and offer directions for future research on the use of social robots for first- and second-language learning.

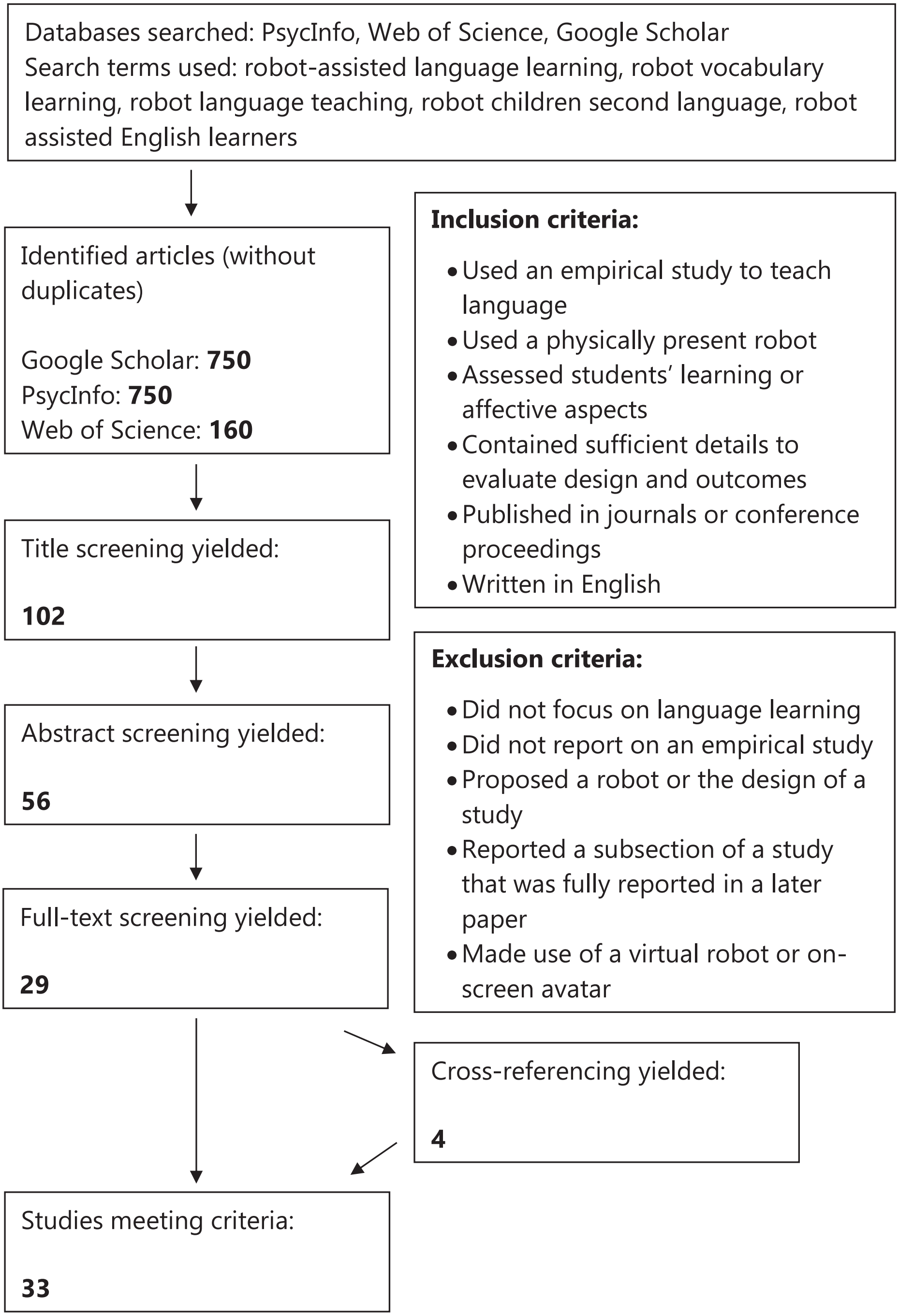
Method

In our review, we take a narrative approach. Specifically, we synthesize the relevant literature in order to provide a comprehensive overview of the work conducted so far (cf. Cronin, Ryan, & Coughlan, 2008). Given the limited number of RALL studies to date, we have adopted an inclusive approach in selecting studies. We did not apply rigorous criteria with respect to the quality of the studies, as due to the emerging nature of the field this could have led to a loss of information (Arksey & O'Malley, 2005).

Figure 1 shows the search, screening, and identification procedure. Studies were included if they: (a) used an empirical design in which language was taught to children or adults (i.e., reviews and studies in which a specific robot or design of a study were proposed were excluded); (b) used a physically present robot (rather than a virtual robot), as we were interested in physical robots that have an embodied presence during the learning task; (c) assessed students' language-learning gains or affective aspects; (d) contained sufficient details to evaluate the design and outcomes (i.e., number of participants, number of target words, learning gains); (e) were published papers in journals or conference proceedings[3]; and (f) were written in English.

Our literature search was conducted using PsycInfo, Web of Science, and Google Scholar. For Google Scholar and PsycInfo, the first 150 results were examined for each search term (cf. Falagas, Pitsouni, Malietzis, & Pappas, 2008; Shultz, 2007). The following five search terms were used: "robot-assisted language learning", "robot vocabulary learning", "robot language teaching", "robot children second language", and "robot assisted English learners". A total of 750 papers in Google Scholar, 750 papers in PsycInfo, and 160 papers in Web of Science were examined based on their

[3] Note that in the field of robotics, many results are presented at conferences rather than in peer-reviewed journals to allow for more rapid development of the field.



Databases searched: PsycInfo, Web of Science, Google Scholar
Search terms used: robot-assisted language learning, robot vocabulary learning, robot language teaching, robot children second language, robot assisted English learners

Identified articles (without duplicates)

Google Scholar: **750**
PsycInfo: **750**
Web of Science: **160**

Title screening yielded:

102

Abstract screening yielded:

56

Full-text screening yielded:

29

Cross-referencing yielded:

4

Studies meeting criteria:

33

Inclusion criteria:

- Used an empirical study to teach language
- Used a physically present robot
- Assessed students' learning or affective aspects
- Contained sufficient details to evaluate design and outcomes
- Published in journals or conference proceedings
- Written in English

Exclusion criteria:

- Did not focus on language learning
- Did not report on an empirical study
- Proposed a robot or the design of a study
- Reported a subsection of a study that was fully reported in a later paper
- Made use of a virtual robot or on-screen avatar

Figure 1. Study selection process.

titles. A total of 102 studies were identified as potentially relevant, as their titles included (parts of) our search terms.

After reading the abstracts of all 102 papers, 46 papers were excluded based on the criteria mentioned above. Specifically, we excluded papers that did not report on an empirical study (*n* = 14; e.g., Belpaeme et al., 2015), did not focus on language learning (*n* = 13; e.g., Arsénio, 2014), proposed a specific robot or a design of a study, rather than an empirical study assessing students' (affective aspects of) learning (*n* = 14; e.g., Funakoshi, Mizumoto, Nagata, & Nakano, 2011), or reported on a part of a study (e.g., preliminary results or a subset of the data), which was fully described in a later published paper that was included in the review (*n* = 5; e.g., Tanaka & Ghosh, 2011).

Subsequently, the full texts of the remaining 56 papers were read, and 27 further papers that did not meet the inclusion criteria were excluded. Reasons for exclusion included proposing a specific robot or a design of a study (*n* = 10; e.g., Nagata, Mizumoto, Funakoshi, & Nakano, 2010), not focusing on language learning (*n* = 6; e.g., Hood, Lemaignan, & Dillenbourg, 2015), reporting on a part of a study only (*n* = 9; e.g., Köse et al., 2015), and the use of a virtual robot rather than a physical one (*n* = 2; e.g., Moriguchi, Kanda, Ishiguro, Shimada, & Itakura, 2011).

Thus, 29 studies met the inclusion criteria. The references of these articles were checked and Google Scholar's "cited by" function was used for each of these articles to identify other potentially relevant studies. In so doing, four additional studies which met the inclusion criteria were found, yielding a total of 33 studies for our review.

Information on the design, characteristics, and main findings were extracted from all 33 studies. Studies were then assigned to one of two categories: language-learning outcomes or affective aspects of RALL. Studies on learning outcomes were grouped according to whether they addressed word learning or other language skills.

Studies on affective aspects were grouped according to whether they focused on motivational aspects, the robot's novelty, or the robot's social behavior. For an overview of all the studies and their characteristics, see Table 1. For an overview of the types of robots used in these studies and their main characteristics, see Table A.1 in the Appendix.

Learning Gains in RALL Studies

Robot-Assisted Word Learning

Word learning in preschool and young school-aged children

Out of all 33 RALL studies included in the review, 13 focused on word learning. Most of these included preschool children or children who just entered school. In three of these, children were presented with words in a second language (L2) or in their first language (L1) over multiple sessions. Pre-tests indicated that the children did not yet know these words prior to the studies, and post-tests indicated that the children learned only few words in each of the three studies.

First, in a study on Japanese child learners of English (L2), an English-speaking Robovie robot was put into several classrooms of six-year-olds and 11-year-olds over a period of two weeks (Kanda, Hirano, Eaton, & Ishiguro, 2004). Children were free in choosing how much to interact with the robot, and could interact with the robot alone or with class mates. Children engaged in various activities with the robot, such as hugging, singing, and playing rock-paper-scissors. The robot used various English sentences, and the authors tested children's knowledge of six different target words and phrases that were commonly used in the interactions between the robot and the children, for example "hello" and "let's play together". The study showed that learning gains were small. On average, the children knew only one or two of the six words or phrases examined in the post-test (Kanda et al., 2004).

Table 1. Methodological Details and Results of Studies Included in this Review

Topic	Study	Participants	Method	Robot	Results
Word learning	Alemi et al. (2014)	N = 46; Age = 12; Country: Iran	Aim: 45 L2 English words; Duration: 5 weeks; 2 conditions: Robot-assisted group and control group	NAO; Autonomous; Role: Teaching assistant	Robot-assisted group learned more words and learned faster than control group
	De Wit et al. (2018)	N = 61; Age = 5; Country: The Netherlands	Aim: 6 L2 English words; Duration: 1 session; 4 conditions: With and without iconic gestures, with and without adaptivity	NAO; Autonomous; Role: Peer tutor;	Gestures benefited retention and engagement; adaptivity benefited engagement
	Eimler et al. (2010)	N = 18; Age = 9 - 11; Country: Germany	Aim: 20 L2 English words; Duration: 1 session; 2 conditions: Robot-assisted group and pencil-and-paper group	Nabaztag; Autonomous; Role: Peer tutor	Robot-assisted group had a more positive learning experience; no significant differences in recall
	Gordon et al. (2016)	N = 18; Age = 3 - 5; Country: United States	Aim: 8 L2 Spanish words; Duration: 7 sessions; 2 conditions: Personalized and non-personalized	Tega; Autonomous; Role: Peer	Children felt more positive towards the personalized robot, although they did not learn more from that robot
	Kanda et al. (2004)	N = 228; Age = 6 & 11; Country: Japan	Aim: 6 L2 English words/phrases; Duration: period of 2 weeks; 1 condition: robot-assisted learning	Robovie; Autonomous; Role: Peer tutor	Most children stopped playing with the robot; only children who continued playing with the robot learned some words
	Kory Westlund et al. (2015)	N = 19; Age = 4 - 6; Country: United States	Aim: 6 L1 English words; Duration: 1 session; 3 conditions: learning with adult, tablet, or robot	Dragonbot; Teleoperated; Role: Tutor	Children learned as much from an adult or a tablet as the robot, but preferred learning with the robot
	Kory Westlund et al. (2017)	N = 45; Age = 5; Country: United States	Aim: 3 L1 English words; Duration: 1 session; 2 conditions: Flat or expressive storytelling	Tega; Teleoperated; Role: Tutor	Children learned a similar amount of words in both conditions, but children were more engaged for the expressive robot than the flat robot
	Kory Westlund et al. (2017)	N = 34; Age = 2 - 5; Country: United States	Aim: 6 L1 English words; Duration: 1 session; 4 conditions: learning with adult or robot (within-subject) and pictures close or far apart (between-subject)	Dragonbot; Teleoperated; Role: Peer	Children could use non-verbal cues from a robot as well as those from a human. Distance did not affect results.
	Mazzoni & Benvenuti (2015)	N = 10; Age = 4 - 6; Country: Italy	Aim: 6 L2 English words; Duration: 1 session; 2 conditions: with child peer or with robot peer	MecWilly; Teleoperated; Role: Peer	Children learned as much with a robot peer as with a child peer
	Meiirbekov et al. (2016)	N = 22; Age = 9 - 10; Country: Kazakhstan	Aim: 10 L2 English words; Duration: 1 session; 2 conditions: ever-winning or ever-losing robot	NAO; Autonomous; Peer	Girls learned most from the ever-winning robot, while boys learned most from the ever-losing robot
	Movellan et al. (2009)	N = 9; Age = 1.5; Country: United States	Aim: 10 L1 words; Duration: 10 sessions; 1 condition: robot-assisted (with 10 control words)	RUBI-4; Autonomously; Role: Tutor	Children learned one word

Table 1. (Continued)

Topic	Study	Participants	Method	Robot	Results
Word learning	Schodde et al. (2017)	N = 40; Age = 24; Country: Germany	Aim: 10 Vimmi (artificial language) words; Duration: 1 session; 2 conditions: adaptive or random	NAO; Autonomous; Role: Tutor	Participants in the adaptive condition did not perform better than those in the random condition, although they improved more during the training
	Tanaka & Matsuzoe (2012)	N = 17; Age = 3 - 6; Country: Japan	Aim: 4 L2 English words; Duration: 1 session; 2 conditions: teaching the robot or not (within-subject)	NAO; Autonomous; Role: Younger peer	Children learned words that they taught to the robot better than the words they did not teach
	Van den Berghe et al. (2018)	N = 67; Age = 4 - 6; Country: The Netherlands	Aim: 6 L2 English words; Duration: 1 session; 3 conditions: learning alone, with child peer, or with robot peer	NAO; Teleoperated; Role: Peer	Children learned as much with a robot peer as with a child peer, although they were outperformed by children learning alone on one task
Reading	Gordon et al. (2015)	N = 46; Age = 3 - 8; Country: United States	Aim: ~ 11 L1 English words; Duration: 1 session; 3 conditions: Inquisitive tablet, inquisitive robot, or non-inquisitive robot	Dragonbot; Autonomous; Role: Younger peer	Children found the inquisitive robot more engaging, but learned more from the non-inquisitive robot
	Hong et al. (2016)	N = 52; Age = 10 - 11; Country: Taiwan	Aim: L2 English skills; Duration: Unclear; 2 conditions: Robot-assisted and control group	Bioloid; Autonomous; Role: Teaching assistant	The robot-assisted group outperformed the control group on reading and listening skills
	Hsiao et al. (2015)	N = 46: Age = 2 - 3; Country: Taiwan	Aim: Emergent L1 literacy skills; Duration: 2 months; 2 conditions: Robot assisted and tablet-assisted group	iRobiQ; Autonomous; Role: Peer tutor	Children in the robot condition improved more on their storytelling ability, word recognition, and story retelling skills than the tablet-assisted children
Grammar	Hyun et al. (2008)	N = 34; Age = 4; Country: Korea	Aim: L1 reading skills; Duration: 4 weeks; 2 conditions: Robot-assisted and tablet-assisted group	iRobiQ; Autonomous; Role: Peer tutor	Children in the robot condition improved more on their story-making, story-understanding, and word-recognition skills than the tablet-assisted children
	Herberg et al. (2015)	N = 23; Age = 10; Country: Singapore	Aim: L2 Latin or French rules; Duration: 2 sessions: 2 conditions: watchful robot or looking-away robot (within subject)	NAO; Autonomous; Role: Tutor	Children performed worse on language tasks when the robot was watching them than when the robot looked away
Speaking	Kennedy et al. (2016)	N = 67; Age = 8; Country: United Kingdom	Aim: L2 French rules; Duration: 1 session; 3 conditions: high verbal availability, low verbal availability, and control	NAO; Teleoperated; Role: Tutor	The robot's verbal availability did not affect learning gains
	In & Han (2015)	N = 16; Age = 10; Country: Korea	Aim: L2 English intonation; Duration: 1 session; 2 conditions: adult or child text-to-speech engine	ROBOSEM; Unclear; Role: Tutor	Children did not learn to vary their intonation more with either text-to-speech engine
	Lee et al. (2011)	N = 21; Age = 10 - 12; Country: Korea	Aim: Speaking and listening skills; Duration: 8 weeks; 1 condition: robot-assisted	Mero & Engkey; Autonomous; Role: Tutor	Children improved their speaking skills but not their listening skills

Table 1. (Continued)

Topic	Study	Participants	Method	Robot	Results
Speaking skills	Rosenthal-von der Putten et al. (2016)	N = 130; Age = 26; Country: Germany	Aim: L2 German skills; Duration: 1 session; 6 conditions: prerecorded speech or robot speech, and physical robot, virtual robot, or no robot	NAO; Unclear; Role: Tutor	Participants did not align lexically or syntactically to the virtual or physical robot
	Wang et al. (2013)	N = 63; Age = 10 - 11; Country: Taiwan	Aim: L2 English skills; Duration: 1 semester; 2 conditions: robot-assisted and traditional class	Tangible learning companions; Unclear; Role: Peer	Children working with the companions had higher motivation, confidence, and engagement
Sign language	Uluer et al. (2015)	N = 42; Age = 21 - 32, 6 - 14, & 10 - 16;	Aim: 10 signs; Duration: 1 session; 2 experiments: one with Robovie and one with the NAO	Robovie R3 & NAO; Teleoperated; Role: Tutor	Experienced sign language learners learned similarly from both robots. Inexperienced learners learned more from the Robovie R3 robot
Motivation	Alemi et al. (2015)	N = 46; Age = 12; Country: Iran	Aim: 45 L2 English words; Duration: 5 weeks; 2 conditions: Robot-assisted group and control group	NAO; Autonomous; Role: Teaching assistant	Students indicated lower anxiety and a more positive attitude towards learning English
	Alemi et al. (2017)	N = 19; Age = 3 - 6; Country: Iran	Aim: L2 English skills; Duration: 10 sessions; 1 condition: Robot-assisted group	NAO; Autonomous; Role: Teaching assistant	Children enjoyed learning from the robot
	Han et al. (2008)	N = 90; Age = 10 - 11; Country: Korea	Aim: L2 English skills; Duration: 1 session; 3 conditions: robot-assisted, non-computer based media, and web-based instruction	IROBI; Autonomous; Role: Tutor	Children showed higher sustained interest and concentration in the robot-assisted group than in the other groups
	Shin & Shin (2015)	N = 66; Age = 14; Country: Korea	Aim: L2 English conversational skills; Duration: 1 session; 2 conditions: robot-assisted and computer-assisted	NAO; Unclear; Role: Tutor	Students in the robot-assisted class participated more and were more satisfied than students in the computer-assisted class
Novelty	Rintjema et al. (2018)	N = 15; Age = 5; Country: The Netherlands	Aim: 17 L2 English words; Duration 3 sessions; 1 condition: Robot-assisted	NAO; Teleoperated; Role: Peer tutor	Children learned most in later sessions, while engagement decreased
	You et al. (2006)	N = 100; Age = 10; Country: Taiwan	Aim: L2 English skills; Duration: 2 weeks; 1 condition: robot-assisted	RoboSapien; Unclear; Role: Tutor	Attention to the robot declined in the second week
Social behavior	De Haas et al. (2017)	N = 59; Age = 3; Country: The Netherlands	Aim: 4 L2 English words; Duration: 1 session; 3 conditions: Adult-like feedback, peer-like feedback, or no feedback	NAO; Teleoperated; Role: Peer tutor	Feedback did not affect engagement, but children worked more independently in the peer-feedback condition
	Saerbeck et al. (2010)	N = 16; Age = 10 - 11; Country: The Netherlands	Aim: Tokipona (artificial language) skills; Duration: 1 session; 2 conditions: social-supportive or neutral robot	iCat; Autonomous; Role: Tutor	Children learning from the socially-supportive robot learned more and were more intrinsically and task-motivated

These outcomes are similar to those obtained in a second RALL study on preschoolers' L2 word learning, by Gordon and colleagues (2016). In this study, a robot that personalized its motivational strategies depending on the child's affective state was used. Specifically, three- to five-year-old English-speaking children played several games on a tablet together with a Tega robot over the course of seven sessions in which they were taught a total of eight L2 (Spanish) words. On average, children learned only one or two out of eight words targeted in this study, as indicated by their scores on a post-test. We will discuss this study's results for personalized motivational strategies further in a later section on the effects of robots' social behaviors.

In the last study on preschoolers' word learning to be reviewed, English-speaking children were taught words in their L1 over multiple sessions (Movellan, Eckhardt, Virnes, & Rodriguez, 2009). Specifically, two-year-old children interacted with a RUBI-4 robot for 10 sessions over a period of 12 days, in which the robot taught the children 10 words through digital and physical games. As in the other two studies, children showed limited learning in this study, as they learned only one or two out of 10 words.

To summarize, limited learning was found in each of the three studies. In all three studies, picture-selection tasks were used to assess children's learning gains. In this type of task, children hear a target word and are asked to choose the picture corresponding to this target word out of several pictures (usually three or four). This task measures receptive vocabulary knowledge, that is, understanding of the meaning of a word. This is in contrast to productive knowledge, or the ability to produce a word with its correct meaning. Crucially, as it is a multiple-choice task, children can also obtain the right answer by guessing, and this chance level should be taken into account when interpreting results. If we do so in interpreting the results of the above studies, it appears that, although the children in Gordon et al. (2016) improved as compared to their pre-test performance, they did not score above chance level for seven out of eight words. Children did score above chance level in Movellan et al. (2009).

Importantly, in the second study, by Kanda et al. (2004), reviewed above, children were free to determine whether and for how long they interacted with the robot, and children's learning was related to the time they had spent interacting with the robot. Interaction time declined for most children already within the first week, and only children who maintained interest during the second week learned some words and phrases. In the studies by Gordon et al. (2016) and Movellan et al. (2009), children's interaction time with the robot was not recorded, making it unclear how much time children actually spent playing with the robot and how active they were during the sessions. One possible explanation of the limited learning gains in these studies, therefore, is that children did not stay engaged enough over multiple RALL sessions to learn a substantial number of words.

Three other RALL studies found more positive results for robots teaching L1 or L2 words to preschoolers. Each of these studies used a different approach and, crucially, consisted of only one session. Specifically, children were taught (a) L1 words through shared book-reading with a robot (Kory Westlund, Jeong, et al., 2017); (b) L2 words by teaching a robot words rather than vice versa (Tanaka & Matsuzoe, 2012); or (c) L2 words by playing "I spy with my little eye" with a robot (de Wit et al., 2018). In all three studies, children learned a substantial number of new words.

In the first study, preschoolers were read one of two versions of a story by a Tega robot (Kory Westlund, Jeong, et al., 2017). The results indicated that, on a receptive vocabulary (picture-selection) task, children responded correctly to two out of three L1 target words on average. Moreover, to measure productive knowledge, a story retell task was used, where children had to retell the story to the experimenter.

Results indicated that children used the target words from the story they had heard more often than those from the story they had not heard.[4]

In the second study that showed considerable word-learning gains in preschoolers, 17 Japanese-speaking preschoolers were taught four L2 (English) verbs by an experimenter who used objects to illustrate the meaning of the verbs (e.g., a cup for the verb "drink"; Tanaka & Matsuzoe, 2012). Then, the child taught the robot two of these words, randomly chosen, by making it act out the relevant verb. The results indicated that children learned the words that they taught the robot better than the words that they did not teach the robot, as evidenced in a picture-selection task. Children demonstrated more knowledge of the verbs that they acted out than those that they did not act out not only in a direct post-test, but also in a post-test three to five weeks after the training.

The last study showing clear word-learning gains in preschool children, conducted by de Wit et al. (2018), tested the effectiveness of a robot's use of gestures in teaching L2. In this study, five-year-old children played the game "I spy with my little eye" with a NAO robot that used an iconic gesture to illustrate the meaning of each target word (e.g., it scratched its head and armpit for the word "monkey") with half of the children, but did not produce such a gesture with the other half of the children. The children's task was to choose a picture of the animal corresponding to the English target word out of several pictures. Immediate post-test results indicated that children learned, on average, almost three out of six words. There was no immediate effect of the robot's iconic-gesture use. However, iconic gestures did benefit retention of the target words: Children who had been presented with iconic gestures in the learning

[4] These data are supported by another study by the same authors, which also showed that children can learn to use words productively through playing storytelling games with a robot (Kory Westlund & Breazeal, 2015). However, only preliminary data from this study have been published so far.

task showed better recall of the words in a delayed post-test one week later than children who had not been presented with iconic gestures.

Overall, these three studies suggest that RALL may benefit children's word learning (de Wit et al., 2018; Kory Westlund, Jeong, et al., 2017; Tanaka & Matsuzoe, 2012). Crucially, all three studies consisted of one session only, suggesting that effects may differ between single- and multiple-session studies. We will return to this issue later in the section on the novelty effect. Some caution is needed, however, in interpreting the results of these studies. An important limitation of the study by Kory Westlund, Jeong, and colleagues (2017) is that children's potential prior knowledge of the words was not assessed. The finding that in this study children recognized not only target words but also control words indicates that they had prior knowledge of these words, as these words were not taught explicitly. This leaves open the possibility that they also had prior knowledge of the target words. A possible limitation of the study by Tanaka and Matsuzoe (2012) is that a human teacher was present in addition to the robot to teach children the L2 words. Since no condition was included in which only a robot was present, we do not know whether children are able to learn from teaching a robot by themselves, or whether the learning in this study was mostly due to being taught by a human adult. Finally, since children are known to learn from teaching someone else (Rohrbeck, Ginsburg-Block, Fantuzzo, & Miller, 2003), it is not clear whether children's word learning was due to the activity of teaching itself (i.e., the additional opportunity to practice the target words), or to teaching a robot specifically. Despite their methodological limitations, the results of these three studies show the potential of using shared book-reading, learning by teaching a robot, and performing language games together with a robot for teaching young children new words in L1 or L2.[5]

An important question is how effective robots are in teaching words in comparison to human teachers. Even though robots are typically not developed with the aim of replacing human teachers, comparisons between robot and human teachers or peers are useful to investigate areas in which robots can complement humans. This question was addressed directly in a study comparing learning gains in an L1 (English) vocabulary-training task provided to preschoolers by a human teacher, a tablet, or a Dragonbot robot (Kory Westlund et al., 2015). Children saw pictures of animals on a tablet and were provided with L1 labels by the human teacher, tablet, or robot. The children in this study learned as much from the tablet or the robot as they learned from the human, that is, four out of six words. Similarly, in a more recent study by the same authors, preschoolers could use non-verbal cues (bodily orientation and eye gaze) of either a human teacher or a robot equally well when mapping unfamiliar L1 (English) words onto pictures (Kory Westlund, Dickens, et al., 2017). Two more studies have investigated how a robot peer compares to a human peer in language-learning experiments. Mazzoni and Benvenuti (2015) found that preschoolers learned as much (i.e., two to three out of six L2 words on average) when working either with a human peer or with a MecWilly robot. Similarly, van den Berghe, van der Ven, and colleagues (2018; Chapter 4 in this dissertation) found that preschoolers generally learned as many L2 words when learning with a child peer or with a robot peer. However, children learning without a peer altogether showed the highest performance. Note that, in this last study, children were provided with L2 words by a human experimenter and played

[5] A pilot study suggests that the learning-by-teaching paradigm could also be effective in improving children's writing skills by having them teach a robot how to write (Hood et al., 2015). However, this finding needs to be investigated further.

games on a tablet with a child peer, robot peer, or without a peer. The presence of the experimenter may have attenuated the possible benefits of a (robot) peer.

This review of studies indicates that (robot) peers do not necessarily lead to higher learning gains than learning without such peers. Rather, the findings of the studies described above suggest that children may be able to learn equally well when being taught by a robot or by a human teacher, or when being assisted by a robot or child peer.

Word learning in school-aged children and adults

As discussed above, RALL studies on word learning in young children show a mixed pattern of results, with some studies reporting small learning gains (Gordon et al., 2016; Kanda et al., 2004; Movellan et al., 2009), and others reporting more substantial learning gains (de Wit et al., 2018; Kory Westlund, Dickens, et al., 2017; Tanaka & Matsuzoe, 2012). Studies with older age groups—older school-aged children and adults—demonstrate a more consistent picture, showing clear word learning across studies. However, very different approaches have been taken across studies, both with respect to the role of the robot (i.e., acting like an assistant vs. a teacher) and whether it was controlled by a human or not, making it difficult to compare results directly.

In a study by Meiirbekov, Balkibekov, Jalankuzov, and Sandygulova (2016), the robot was used as a peer learner. Children’s task in this study was to play a game together with a NAO robot in which they were provided with a letter and had to select images of words starting with that letter. After one lesson, children were on average able to produce over three out of the 10 L2 words that they had been taught.

In contrast, in another study (Alemi et al., 2014), the robot was used as a teaching assistant. Here, a NAO robot assisted in teaching young adolescents L2 (English) words by interacting with the students, making gestures depicting the target words, showing pictures, and telling stories. Students were taught a total of 45 words over the course of 10 sessions. The classes incorporating the robot were compared to an English class that did not have a robot assistant but engaged in the same type of activities. Results indicated that the students in the RALL classes learned faster, learned more, and retained more words than the students educated in the traditional class.[6]

Yet another study (Eimler, von der Pütten, & Schächtle, 2010) had 9- to 11-year-old German children play L2 English games with a Nabaztag robot for one session. The results indicated that children learned almost 14 out of 20 words on average. These are very high learning gains. Crucially, however, these learning gains did not significantly differ from those of children who had been taught these words through paper vocabulary lists. This suggests that children of this age may generally be skilled word learners and obtain high learning gains across different types of vocabulary interventions.

Finally, a study on adults learning words in an artificial language used the robot as a teacher (Schodde, Bergmann, & Kopp, 2017). The participants in this study were taught 10 words in the artificial language Vimmi via an "I spy with my little eye" game. In each trial, a NAO robot asked the participant to find the picture of the target word among distractor pictures. Participants' knowledge of the target words was assessed in an immediate post-test via two translation tasks: one from Vimmi to German and one from German to Vimmi. Participants produced, on average, seven out of 10 words in the Vimmi-to-German translation task and 3.5 out of 10 words in the German-to-Vimmi translation task. These learning gains are substantial, especially given that (a) translating words is more difficult than a receptive task (Mondria & Wiersma, 2004);

[6] A pilot study with four autistic children and a design similar to that in Alemi et al. (2014) suggests that robot-assisted language classes are also effective for autistic children (Alemi, Meghdari, Mahboub Basiri, & Taheri, 2015), although further research is required on this subject.

(b) there was only one session; and (c) the learning task consisted of only three trials per target word.

Apart from the different roles assigned to the robot, another aspect that makes RALL studies on word learning difficult to compare is that, in many of the studies described above, the robot was teleoperated by the experimenters (see Table 1 for information on whether studies used a teleoperated or autonomous robot). Teleoperation refers to a person controlling the robot, often without the participant's knowledge, in real-time. Teleoperation is often the preferred or even the only option for certain tasks, as an autonomous robot (working without teleoperation) would require speech recognition and predefined scripts. Such scripts describe all the steps of an interaction, and the robot cannot divert from this script. Elaborate scripts are needed to have robots respond appropriately to the input, but even then the suitability of the responses cannot be guaranteed due to, amongst other reasons, the unpredictability of participants' behavior. Thus, previous studies that used an autonomous robot typically consisted of simple designs (such as "I spy with my little eye" games on a tablet) that allow for limited variability in the learner's responses (de Wit et al., 2018; Schodde et al., 2017). In more complex settings, experimenters can ensure through teleoperating that the robot answers appropriately, as they can simply type in contingent responses. Hence, given the current state of robot technology and the scientific literature, how effective robots are when operating autonomously remains an open question.

Summarizing RALL studies on word learning across age groups

The L1 and L2 word-learning studies discussed above found mixed results regarding the robot's effectiveness for word learning. Specifically, several studies found only small (Movellan et al., 2009) or no learning gains (Gordon et al., 2016), or learning gains that only held for a subgroup of the children (Kanda et al., 2004). Other small-scale studies with preschool children showed positive effects of the use of robots in word learning and suggest that aspects such as learning by teaching and gestures might improve learning gains (de Wit et al., 2018; Tanaka & Matsuzoe, 2012). However, many studies were based on small samples and/or lacked control conditions and therefore provide only tentative evidence. Studies on school-aged children (Alemi et al., 2014; Eimler et al., 2010; Meiirbekov et al., 2016) and adults (Schodde et al., 2017) suggest that RALL benefits word learning more with these groups than with preschool children. However, direct comparisons between adults and children are needed to support this conclusion. Furthermore, it is important to note that most of the studies showing high word-learning gains employed the robot as a teaching assistant or peer learner rather than as an independent tutor (Alemi et al., 2014; Meiirbekov et al., 2016; Tanaka & Matsuzoe, 2012). Perhaps, in their current form, robots are not sufficiently technologically advanced (e.g., due to difficulties with speech recognition) to be effective tutors on their own. The current evidence base suggest that teleoperation is still required for robots to be effective tutors and that technological advances and research on which robot behaviors are effective for learning are required to develop effective autonomous robot tutors.

Language Skills Beyond Word Learning

Language use comprises more skills than just vocabulary. These other skills, such as reading, speaking, grammatical skills, and sign language, have been studied less extensively in RALL research than word learning; only 11 of the 33 selected studies addressed (one of) these skills.

Reading skills

RALL studies on reading skills show that a robot may be beneficial in assisting the teaching of reading skills, either in the function of an assistant or as a tutor. Specifically, comparing an L1 robot-assisted digital book-reading program to the same program without a robot, Hyun, Kim, Jang, and Park (2008) found that preschoolers in the robot-assisted program improved more on story-making, story-understanding, and word-recognition skills over a four-week period than children who were not assisted by the robot. Similar results were obtained in another study on early L1 reading (Hsiao, Chang, Lin, & Hsu, 2012). In this study, two-year-old children followed an early L1 reading program over a period of two months, either supported by an iRobiQ robot with a screen or by a tablet without a robot. The results indicated that both groups improved on early literacy tests measuring comprehension, storytelling ability, retelling of stories, and word recognition. However, the children who had interacted with the robot improved more on their storytelling ability, word recognition, and story-retelling skills than children who had worked with a tablet only.

While the results of these two studies are promising, a third study on L1 reading in young children did not find such positive results. In this study, a relatively large group of 46 preschoolers performed a learning task in which they had to find out, together with a Dragonbot robot, how to read words (Gordon, Breazeal, & Engel, 2015). On average, the children learned the written word form of only one out of eleven new words. As in some of the other word-learning studies reviewed above (Gordon et al., 2016; Kanda et al., 2004; Movellan et al., 2009), these small learning gains were taken as evidence of the robot's effectiveness by the authors.

The only RALL study on L2 reading that has been performed to date, found a positive effect of the presence of a robot teaching assistant on children's L2 (English) reading skills (Hong, Huang, Hsu, & Shen, 2016). In this study, either a human or robot teacher taught 10- to 11-year-old children reading, speaking, and listening skills by reading stories aloud, encouraging children to read sentences out loud, engaging in act-out games, and engaging in question-answer conversations. Children in the robot-assisted classroom outperformed children in the traditional classroom on a standardized English reading test. Children in the robot-assisted classroom were highly motivated by the robot, which may have benefited their learning as compared to children in the traditional classroom. Overall, these findings suggest that there is potential for robots supporting reading skills.

Grammar

Two RALL studies addressed L2 grammar learning and both demonstrated positive effects of the robot on children's learning. First, Kennedy and colleagues (2016) found that a NAO robot positively affected English-speaking children's learning of the French articles 'le' and 'la'. The robot tutor taught eight- to nine-year-old children three rules regarding French articles. The children improved their knowledge of French articles and retained this knowledge in a post-test one week later. In the second RALL study on L2 grammar learning, Herberg, Feller, Yengin, and Saerbeck (2015) investigated children's learning of Latin and French rules, such as those governing plural and article use, in two separate sessions with a NAO robot. The robot either looked at them or looked away during tasks in which the children had to practice the newly acquired information. The study showed that children learned the rules from the robot. Unexpectedly, however, children performed worse if the robot had looked at them, although the effect was found for difficult items in Latin only. A possible explanation of this finding, proposed by the authors, is that instead of representing a comforting social presence during the task and putting the child at ease (which was the intended outcome), the robot increased pressure and, as such, made the children perform worse. These results indicate not only that the specific learning materials and their difficulty may affect experiment outcomes, but also that the robot's behavior may affect learning in unexpected ways.

Speaking skills

Studies addressing L2 speaking skills found mixed results. One study used a ROBOSEM robot to teach Korean-speaking children to use English intonation patterns (In & Han, 2015). Native English speakers vary their intonation more than native speakers of Korean, and less varied intonation shows Korean L2 English learners' non-nativeness. In the study by In and Han (2015), children did not learn to vary their English intonation upon interacting with the robot as compared to their pre-test performance. The experimenters concluded that the robot's speech system (as opposed to human speech) is not effective enough to evoke changes in intonation. However, another study, also conducted in Korea and aimed at improving L2 English speaking and listening skills, did find improvement in other speaking skills (Lee et al., 2011). Specifically, this study examined children while they were playing with two robots, the Mero robot and the Engkey robot, with the purpose of improving their L2 (English) speaking and listening skills. The study showed that children's L2 listening skills did not improve upon interacting with the robots, but that L2 speaking skills (measured through pronunciation, vocabulary, grammar, and communicative ability) did improve (Lee et al., 2011). Interestingly, the children in this study improved on all four aspects of speaking skills.

Even though both studies involved the same L1 and L2, they show opposing results, as the participants in Lee et al. (2011) improved their L2 pronunciation upon interacting with the robot, whereas the children in the study by In and Han (2015) did not. Contradictory results were also found in two studies that compared robot-assisted classrooms to traditional classrooms in teaching L2 English speaking skills to Taiwanese children: In one study, children improved their speaking skills more in the robot-assisted classroom than children in traditional classrooms (Wang, Young, & Jang, 2013), while in another study, children in a robot-assisted classroom outperformed students in a traditional classroom on L2 listening, but not on L2 speaking (Hong et al., 2016). This contrast in results may be due to the different scope of the L2 training: The training of Wang et al. (2013) was only aimed at teaching speaking skills, while the training of Hong et al. (2016) was also aimed at teaching listening, reading, and writing.

Conflicting results across studies targeting the same skill (i.e., L2 speaking skills) in very similar participant groups may be due to the various ways in which speaking skills were evaluated. Speaking skills can be assessed in different ways, for example by measuring intonation, speech rate, pronunciation, vocabulary, and grammatical complexity. Given the very different operationalizations of (L2) speaking skills in earlier work, future work on RALL assessing these different aspects would be useful to identify which speaking skills benefit most from robot tutoring. In pursuing this line of research, an important recommendation is that studies target the same L1s and L2s to test the effectiveness of robots for teaching speaking skills, as most L2 speaking skills are heavily dependent on learners' L1.

Before we conclude this section on RALL research on L2 speaking, a final study is noteworthy, in which adults' L2 lexical and syntactic alignment behavior was assessed. Lexical and syntactic alignment refers to the degree to which speakers adapt their words and sentence structures to those of their conversational partner (Rosenthal-von der Pütten, Straßmann, & Krämer, 2016) and thus involves a very different type of learning (implicit vs. explicit) and skill than the type of speaking skills (e.g., pronunciation and intonation) discussed above. Rosenthal-von der Pütten and colleagues compared the L2 (German) alignment behavior of adults with various L1s to a physical robot, a virtual robot, and a computer system with pre-recorded speech without a (virtual) agent. Contrary to the authors' expectations, participants showed no alignment to the physical or virtual NAO robot or the computer system (i.e., they did not use similar words and sentence structures). Furthermore, there were no significant differences in the perceived human-likeness of the robot's text-to-speech system (i.e., the system that converts text into spoken voice output) and the pre-recorded human speech. This is a striking result, as text-to-speech systems are often of inferior quality to human speech. It may also explain the absence of alignment effects: Participants may not have felt the need to align to a computer with such an advanced speech system. Note that alignment may also not result in implicit learning if the speech system is perceived to be of inferior quality: Learners may not learn advanced vocabulary or grammar from inferior systems. Clearly, more research is needed on how a robot's text-to-speech system affects L2 learning in general and the learner's pronunciation of L2 words in particular.

Sign language

RALL studies on sign language are nearly absent, and only one out of the 33 in our review addressed this topic. In this study, robots were found to be able to teach sign language successfully to various types of learners (Uluer, Akalın, & Köse, 2015). Uluer and colleagues compared the effectiveness of two robots in teaching Turkish sign language to three groups of Turkish participants: hearing adults, hearing children and hearing-impaired children. The first robot, a Robovie R3 robot, has hands with five independent fingers, allowing for the production of signs that are more accurate than those of most other robots. The second, a NAO robot, has only three fingers that cannot be moved independently. The three participant groups played imitation and act-out games with the robots. The results indicated that all groups learned most of the signs from the robot. Even though there was no difference between the effects of the two robots for the experienced sign language learners, the Robovie R3 robot resulted in significantly higher learning gains than the NAO robot in the inexperienced learners (typically hearing groups, who, unlike the hearing-impaired children, were novices in sign language). Thus, considering the specific characteristics of the robot seems especially relevant in learning situations like these, which rely more on the robot's physical interaction possibilities.

Summary

Studies on language skills other than word learning are rare in RALL research. Also, they are typically diverse, in the sense that they have looked at different age groups and used very different research designs. The available studies, albeit few in number, suggest that a robot can successfully assist in teaching reading skills (Gordon et al., 2015; Hong et al., 2016; Hsiao et al., 2012; Hyun et al., 2008), grammar learning (Herberg et al., 2015; Kennedy et al., 2016), and sign language (Uluer et al., 2015), either in L1 or L2. The evidence with respect to speaking skills is more mixed (Hong et al., 2016; In & Han, 2015; Lee et al., 2011; Rosenthal-von der Putten et al., 2016; Wang et al., 2013), and may differ depending on which types of speaking skills are assessed (e.g., pronunciation, intonation, lexical alignment; cf. In & Han, 2015; Lee et al., 2011; Wang et al., 2013 for pronunciation; Rosenthal-von der Putten et al., 2016 for alignment).

Affective Aspects of RALL

Robots' Positive Effects on Motivation

Robots do not only affect language-learning gains, but may also affect students' learning strategies and motivation to learn. Given that motivation has been found to be positively related to learning achievements (Dörnyei, 1994; Pekrun, Goetz, Titz, & Perry, 2002), it is important to look at how the use of robots in language-learning studies affects students' motivation. Previous studies indicate that robots generally have a positive effect on students' motivation in RALL.

A number of studies comparing a robot-assisted classroom to a traditional classroom found higher student motivation in robot-assisted classrooms than in traditional classrooms. In the study by Alemi et al. (2014) on L2 word learning in school-aged students that was reviewed above, robot-assisted students indicated that they felt very positive about learning with a robot. As discussed earlier, learning outcomes in this study indicated that the robot-assisted students learned faster, learned more, and retained more than the students in the traditional class. In fact, the students in the robot-assisted class needed less than a third of the time required by the traditional class to work through the materials.

The effects of RALL on these students' learning-related emotions were reported in a follow-up paper (Alemi, Meghdari, & Ghazisaedy, 2015). Using self-report questionnaires, the authors found that students were less anxious to make mistakes and less self-conscious about making mistakes in the presence of the robot than in the presence of a human teacher. Similar positive effects were found in studies on speaking skills. Ten- to 11-year-old students in Wang et al. (2013) who were taught together with a robot companion also displayed higher confidence, motivation, and engagement than children in a traditional classroom. A positive effect on students' motivation was also found by Lee et al. (2011), who observed that a robot improved learners' self-reported satisfaction, interest, confidence, and motivation. Finally, the nine- to 11-year-old children in the study by Eimler et al. (2010) were found to have a more positive learning experience when being taught L2 English words with assistance from a robot than when they were taught these words through paper vocabulary lists, even though there were no significant differences in word learning between the two conditions.

Other studies have compared the motivational aspect of the robot to other types of technology. The preschool children in Hsiao et al.'s (2012) reading experiment participated much more actively when assisted by a robot: They engaged more in reading, singing, and replying to questions than when working without robot. An observational study found that preschoolers in an L2 (English) learning class showed less anxiety, higher motivation, and higher engagement after interacting with a robot multiple times (Alemi, Meghdari, & Sadat Haeri, 2017). Furthermore, in a study comparing the at-home use of the IROBI robot for L2 (English) language learning to non-computer based media and web-based instruction, 10- to 12-year-old children working with a robot showed longer sustained interest and concentration than the other groups (Han, Jo, Jones, & Jo, 2008). Similarly, 14-year-old students were found to participate more and to be more satisfied when working with a NAO robot in an L2 (English) conversation class than when working with a computer (Shin & Shin, 2015). The students' motivation did not differ across conditions. These results must be interpreted with caution, however, as the students working with the robot engaged in an additional group conversation with the robot and thus had more exposure to the technology. Last, in Kory Westlund et al.'s (2015) study, preschoolers' learning with a robot was compared to learning with a tablet and a human teacher. The authors found that almost all the children preferred being taught by the robot to being taught by the human teacher or the tablet. Note, however, that this preference did not lead to higher learning outcomes. In summary, robots seem to have a more positive effect on students' motivation than other types of technology, such as tablets or web-based programs.

The positive effects of robots on learning-related emotions have not only been found in RALL studies, but also in studies looking at other types of robot-assisted learning, such as programming and drawing and interpreting graphs (Chin, Hong, and

Chen, 2014; Lee & Lee, 2008; Liu, Lin, & Chang, 2010; Mitnik, Recabarren, Nussbaum, & Soto, 2014; Nourbakhsh et al., 2005). Interestingly, the picture that emerges from the literature on affective aspects of robot-assisted learning is much clearer than that on language-learning gains: The assistance and/or presence of social robots has a positive effect on students' engagement, attitude and motivation, and this holds across domains (language vs. other domains) and across age groups. This suggests that the potential of robots lies mainly in their ability to motivate students.

Interestingly, such positive effects on motivational aspects are generally not found for other types of technology, such as interactive white boards, blogs, and virtual worlds, for which only weak evidence of positive effects is found (Barrett & Liu, 2016; Golonka et al., 2014). It should, however, be studied further, as people are likely to differ in the degree to which they feel intrinsically motivated to make use of technology for language learning (Stockwell, 2008). Future research could investigate the degree to which students are intrinsically motivated to work with robots and whether and how the positive effects of robots on motivation could benefit students' language learning. One caveat is noteworthy here. It is not completely clear to which extent the boost in motivation is due to the motivational actions of the robot itself or by the novelty of the robot. On the basis of the current state of knowledge, the possibility remains that the robot initially boosts motivation, but that this effect fades out over time as people become accustomed to working with robots. This possibility will be discussed further in the next section.

The Novelty Effect in RALL research

Robots often spark a lot of enthusiasm in their users. This excitement can result in a so-called novelty effect on learning: Learners enjoy the new technology so much that their initial interest leads to higher learning outcomes, which would not have been attained if learners had been more familiar with the robot (cf. Liu, Liao, & Pratt, 2009).

Once learners become used to the technology, their interest and boost in learning outcomes fade away. This effect might be particularly influential in experiments involving one session or a small number of sessions. In fact, it may, at least in part, explain why one-session word-learning studies found higher learning outcomes than word-learning studies consisting of multiple sessions.

Many authors do not report on how novelty may have affected their results or on how they controlled for a novelty effect. Some one-session experiments have addressed the issue of the novelty effect by having the children play with the robot for a few days before the actual experiment (e.g., Han et al., 2008). It is not clear, however, whether this procedure attenuated the novelty effect in this study, as the experiment itself consisted of only one session.

Studies reporting on students' interest in robots over time found mixed results. In the previously discussed study on L2 word learning by Kanda et al. (2004), the amount of time in which children wanted to interact with the robot quickly decreased within two weeks, and this decrease in interaction time with the robot in turn attenuated the learning effect. Specifically, in this study, learning gains were only found for the children who continued playing with the robot, a subset of about a quarter of the 200 participants in the study. Moreover, the continued interaction was not due to sustained interest. Rather, most children indicated that they continued playing with the robot out of pity (Kanda et al., 2004).

A similar decline in interest in working with the robot is reported in a study in which a RoboSapien robot assisted a teacher in English classes, engaging in several activities such as storytelling, answering questions, cheerleading, gesture games, and pronunciation exercises (You, Shen, Chang, Liu, & Chen, 2006). Overall, the children enjoyed the robot, although the attention they paid to the robot declined in the second week. Already after two lessons, children had gotten used to the robot and became less interested in working with it. Language-learning gains were not assessed, so it is not clear whether the decline in children’s motivation affected learning.

Similarly, a decline in task engagement was found over the course of three sessions in a study on preschoolers learning L2 English words with a NAO robot (Rintjema, van den Berghe, Kessels, de Wit, & Vogt, 2018). This decline in task engagement did not impact learning gains, as children learned more in later sessions than in the first session. These results need to be interpreted with caution, as the specific target words taught in the lessons and the type of the lessons were not counterbalanced. Therefore, it is not clear whether changes in engagement and learning were due to a (dissipating) novelty effect of the robot or to differences in the content of the lessons. However, the studies discussed above do show that further development of both technology and content are needed to sustain children’s interest and to make children enjoy interactions over time in order for robots to become effective learning companions or tutors.

In contrast to the studies summarized thus far that showed a decline in participants’ interest in the robot, two previous studies found that participants sustained interest in working with a robot over a longer period. In the first, by Alemi, Meghdari, and Ghazisaedy (2015), a relatively large sample of students reported positive experiences after having worked with a robot for 10 sessions over five weeks, suggesting that they maintained their interest in the robot over multiple sessions. A possible explanation is that the robot functioned as an assistant to a human teacher, and that the teacher and robot together could sustain students’ interest for a prolonged time. If a robot is solely responsible for maintaining an interaction, the behavioral and conversational demands on the robot’s social interactional skills are higher than if a human teacher can act as a mediator.

The second study showing sustained interest found that the robot could maintain children's attention even if it interacted with the child independently, at least in very young children (Hsiao et al., 2012). The toddlers in this study interacted with a robot twice a week for a period of eight weeks. Children did not lose interest in the robot and participated equally actively in the last four weeks as in the first four weeks. Note, however, that children in the control condition who worked with a tablet also sustained their interest over this period. This suggests that the e-book that was used as teaching material in both conditions, which contained many different activities, was interesting enough to sustain interest over a long time period.

Raising and maintaining participants' interest is crucial to successful interactions, and recent work has addressed the issue of maintaining interest in RALL situations. Specifically, Han, Kang, Park and Hong (2012) conducted several pilot RALL studies with a IROBIQ robot, and concluded that there are several strategies to encourage sustained interaction between a robot and children. These strategies are mostly focused on making the child seem important to the robot. This can be achieved by having voice- or face-tracking systems recognize and track the child, using pictures of the child on the screen, "remembering" the child's learning history, or working around quirks (e.g., framing quirks by telling the robot's "birth story"; Han et al., 2012). Therefore, a key recommendation for future RALL studies, according to these authors, is to teleoperate the robot in order to tailor the robot's speech and actions to the specific behavior and needs of an individual child. Currently, artificial intelligence, visual recognition systems, and automatic speech recognition systems clearly do not yet allow for robots to interact autonomously with a child in such a way that the child will remain interested in the robot.

This recommendation is in keeping with the conclusion of a review of several (mostly non-RALL) robot studies in which robots interacted with children and adults over longer periods of time (Leite, Martinho, & Paiva, 2013). These studies found varying results, from a clear, short-lived novelty effect (Fernaeus, Håkansson, Jacobsson, & Ljungblad, 2010) to sustained interest over a period of five months (Tanaka, Cicourel, & Movellan, 2007). The authors propose several guidelines to encourage long-term interaction with social robots, involving the robot's appearance, behaviors, affect, memory, and adaptation. For example, one recommendation is that the robot should have both routine behaviors, such as greetings, as well as new and personalized behaviors over time (e.g., adding new games or suggesting games that match participants' interests). It is likely that the effectiveness of behaviors aimed to increase or sustain learners' interest differs per target group (e.g., depending on age, gender, or subject), and a robot's behaviors should be focused on its audience.

In short, the novelty effect is an important issue to be taken into account in robot studies. At least some results on learning-related emotions and learning gains obtained in previous robot studies are likely to stem from the initial excitement when learners work with a robot for the first time. Some ways in which long-term interaction could be fostered involve working around technical limitations (e.g., teleoperating the robot) or increasing (diversity in) the robot's social behavior. The next section will outline in more detail the outcomes, and concomitant complexities, of earlier work on robots' social behavior, and in particular on their supportiveness and motivational behavior.

The Complexity of Robots' Social Behavior

As noted in the Introduction, one of the advantages of robots is their appearance, and therefore their potential benefits in establishing more natural interactions. Robots can be programmed to behave socially via both non-verbal behaviors (e.g., gaze, body posture) and verbal behaviors (e.g., giving praise, saying someone's name). This section reviews evidence relating to robot behavior's ability to increase students' motivation and learning outcomes.

Several studies have examined how a robot's social behavior may positively affect learning gains. In one of these, the effect of social support on children's ability to learn an artificial language was investigated. Specifically, Saerbeck, Schut, Bartneck, and Janse (2010) studied how ten- to 11-year-old children interacted with the iCat to learn an artificial language. The experimenters manipulated the degree to which the robot was socially supportive, such that in one condition the robot engaged in a social dialogue, while in the other condition the robot focused solely on the desired transfer of knowledge. Children interacted with the robot for equally long periods across the two conditions, but children working with the socially-supportive robot learned more and were more intrinsically motivated, as they reported having had more fun than children working with the neutral robot. This finding is similar to that of Gordon et al. (2016), who found that children felt more positive towards a personalized robot. Crucially, in this study, the robot adapted its motivational utterances to the child's affective state (e.g., excited, thinking, or frustrated). Note, however, that this did not lead to higher learning gains, in contrast to the results of a non-RALL study by Leyzberg, Spaulding, and Scassellati (2014), in which a personalized robot tutor resulted in higher learning outcomes than a non-personalized tutor in a puzzle-solving task. Such mixed findings indicate that personalization is an important avenue for future research on exactly how robots can be used as effective tutors. Note that the two studies used very different age groups (preschoolers and adults), and personalization may affect age groups differently.

Another type of robot social behavior that may benefit learning is the robot's expressiveness (Kory Westlund, Jeong, et al., 2017). Using storytelling to teach preschoolers L1 (English) words, these researchers compared an expressive robot to a ‘flat’ robot. The expressive robot spoke in a more expressive way, with changes in its intonation. The ‘flat’ robot did not vary the intonation of its utterances. Both voices were recorded by a female adult rather than created via text-to-speech systems, as a computer-generated voice cannot reach the same variation in intonation as a human voice. The expressiveness of the robot did not affect how many target words children recognized or how they perceived the robot. However, children in the expressive condition used more target words in their retellings, told longer stories in a delayed post-test four to six weeks later, and were more likely to imitate the robot’s phrasing in their story retellings. Crucially, concentration, engagement, and surprise (but not attention during the story) were significantly higher for children in the expressive condition than in the flat condition. Thus, although children did not learn more words receptively when interacting with an expressive robot, the expressiveness of the robot had a positive effect on the way in which children were involved in the task and on their production and retention of the target words.

Other aspects of social behavior, however, do not seem to have such positive effects on language learning. Specifically, previous work on the effects of verbal availability, feedback, and adaptivity has shown mixed results. The term “verbal availability” refers to a robot’s sensitive response towards a student, for example by using the student’s name, giving praise, or asking for the student’s opinion (Kennedy et al., 2016). In the study by Kennedy and colleagues (2016) that was also discussed above, a NAO robot taught eight- and nine-year-old children French articles, showing either high or low verbal availability. High verbal availability did not result in greater learning gains. Interestingly, however, another study by the same authors, reporting on a math-learning experiment, found that the NAO’s nonverbal availability (i.e., the use of nonverbal cues such as gaze and posture to attend to the student) did affect children’s learning gains positively (Kennedy, Baxter, Senft, & Belpaeme, 2015), indicating that the effects of verbal and non-verbal availability may defer depending on the specifics of the tasks used.

Similarly, the effect of a robot's feedback during RALL tasks is unclear. To date, only one study has directly compared the effects of several types of feedback in a RALL task. In this task, three-year-old children were taught L2 English count words by a NAO robot (de Haas, Baxter, de Jong, Krahmer, & Vogt, 2017). There were three conditions in which children were given: (a) explicit positive and implicit negative feedback (i.e., adult-like feedback); (b) explicit negative feedback (i.e., peer-like feedback); or (c) no feedback. The authors did not assess children's vocabulary learning gains, but studied how feedback affected children's engagement, as measured via eye gaze and the amount of help children needed from the experimenter. The study showed that children looked more often at the experimenter in the no-feedback condition than in both feedback conditions, and that they needed more help from the experimenter in the explicit positive and no-feedback conditions than in the explicit negative feedback condition. There were no differences in the duration of the gaze towards the stimulus materials and the robot across the three conditions. This study indicates that the way in which children engage in a learning task is affected by the feedback the robot provides, but more research is needed to assess how a robot's feedback affects learning gains and motivation, and ideally to compare how children respond to feedback from robot and human tutors.

A final behavior that may affect learning is adaptivity. This is an area worth exploring, since robots can, at least in theory, be programmed to provide adaptive tutoring. Only two studies to date have studied the effects of robot adaptivity on language learning. In the first study by Schodde et al. (2017), an adaptive robot system was compared to a random system in teaching German adults words from an artificial language called Vimmi. The adaptive robot system selected which item to teach (depending on the items that the participant showed difficulty with) and the difficulty of the item (i.e., the number of distractors). The adaptive robot did not result in higher scores on two translation tasks (from Vimmi to German and from German to Vimmi) than the random robot, but participants in the adaptive condition improved more within the “I spy with my little eye” game (i.e., they found the right target more often) than participants in the random control condition. Schodde and colleagues note that the fact that the participants’ greater improvement did not result in higher learning gains could be due to the difficulty of the translation tasks as compared to the leaning task. If the participants’ knowledge had been measured receptively, a benefit of adaptivity might have been found.

However, in the second study examining the effects of a robot’s adaptivity on language learning (de Wit et al., 2018), no positive effects of adaptation were found on word-knowledge tasks either. In this study, which is also discussed above, de Wit et al. (2018) investigated the effect of adaptivity on Dutch preschoolers’ learning of L2 English animal names. For half of the children, the “I spy with my little eye” game was adapted to the child’s needs (e.g., fewer distractor pictures for difficult target words), and for the other half of the children, the difficulty was not adapted. While children in the adaptivity condition remained engaged during the game, in contrast to the children in the no-adaptivity condition, adaptivity did not result in higher learning gains. As these studies do not convincingly show that adaptivity results in higher learning gains, more research is needed to study the effect of adaptive systems and to confirm the importance of adaptivity in RALL.

Thus, across studies, there are contradictory results with respect to the effects of the robot’s personality and social behavior on learners’ motivation and learning outcomes. Although one could adopt a ‘no harm in trying’ policy with regard to incorporating personalized or social behavior in child-robot interactions, other experiments indicate that caution is needed, as social behavior does not always lead to higher learning gains. In the study by Gordon et al. (2015), for example, a robot that was not curious led to higher learning gains than a curious robot that showed excitement about the learning task and interest in the progression of the story. A possible explanation for this finding is that the curious robot was indeed more engaging, but distracted participants from the learning task and therefore resulted in smaller learning gains compared to a less engaging robot.

These results are reminiscent of the study by Herberg et al. (2015) described earlier, which showed that children performed worse when a robot looked at them during tasks than when it looked away. Finally, support for the idea that social behavior may result in lower learning gains comes from an experiment in which prime numbers were taught by a NAO robot to seven- and eight-year-old children (Kennedy, Baxter, & Belpaeme, 2015b). In this study, an anti-social robot (which actively avoided gaze) resulted in greater learning gains than a social robot. In-depth analyses revealed that children spent more time looking at the social robot than the anti-social robot, thus looking less at the educational content provided by the tablet. These findings suggest that when a robot is too social, it can distract the child and actually make the child learn less than when the robot is less social, at least when the educational content is provided by an external medium such as a tablet and not by the robot itself.

A finding that adds to the complexity of the effects of a robot's social behavior on task interest and learning is that gender can play a role in the beneficial or adverse effects of the robot's social behavior. As discussed above, the children in Meiirbekov et al.'s (2016) study learned to produce over three new L2 words when working with a robot. The experimenters compared learning gains for children working together with a robot that would always either win or lose the game to assess whether the robot making mistakes would make the child feel more at ease during the learning process.

Interestingly, the child’s gender determined which robot version led to the highest outcomes: Girls learned twice as many words as boys from the ever-winning robot, whereas boys learned twice as many words as girls from the ever-losing robot. The experimenters did not offer any possible explanations for this interaction effect of robot condition and gender, but it may suggest that there are differences between girls and boys in empathy or competitiveness, that is, in how they perceive the different robot versions (e.g., they may feel sorry for the robot when it loses or focus on their own wins) and in how they engage with the robot when it always either wins or loses.

To summarize, the existing evidence with respect to the robot’s social behavior is mixed. A robot showing social behavior such as producing the child’s name can increase children’s engagement in learning tasks. At the same time, the social behaviors of the robot can distract children from learning and, as a consequence, result in poorer learning outcomes. Moreover, there may be interaction effects with child characteristics such as gender, and results may differ depending on learners’ sociocultural backgrounds. The studies listed above have been conducted in countries all over the world (e.g., the US, the UK, Singapore), and the different contexts may affect how children respond to the robot’s behavior. Moreover, the studies reviewed in this section involved a single session only, and it is not clear whether effects of robots’ social behavior differ when learners interact with robots over multiple sessions. Thus, it is still difficult to disentangle the effects of the robot’s social behavior shown from the novelty effect discussed above. Future research should try to optimize the social behavior of the robot for different learning tasks (e.g., grammar learning vs. speaking skills) and different groups of learners (e.g., preschool vs. school-aged children, girls vs. boys) and incorporate adaptivity and feedback.

Discussion

The goal of this review was to provide an overview of the current evidence on RALL and to identify potential topics for future research regarding the use of robots for language teaching. Thirty-three studies addressing word learning, reading skills, grammar, speaking skills, and sign language have been discussed, focusing on two important aspects: (a) the robot's effect on children's L1 and L2 learning gains and learning motivation; and (b) the way robots should behave to maximize learning outcomes. Below, these aspects will be discussed separately, followed by a discussion of possible avenues for future research.

Mixed results were found with respect to L1 and L2 learning outcomes. Most studies focused on word learning, and did not clearly show whether robots are effective for word learning. More research is needed to determine the most effective role for the robot (e.g., teaching assistant or peer learner), the age groups for which robots are most beneficial (e.g., preschool children, school-aged children, or adults), and the optimal number of sessions for word learning (one or more). The few studies examining reading skills, grammar learning, and sign language showed quite positive results, while the evidence with respect to speaking skills is more mixed. Note that the studies made different comparisons: Studies on grammar learning and sign language compared different robot behaviors or platforms to assess the most effective robot (behavior), while the studies on reading and speaking skills compared the effectiveness of a robot to other types of technology or traditional classrooms. Moreover, the conflicting results between skills may result from differences in demands on the robot's interactional qualities (such as being able to have contingent conversations), which are likely higher in lessons on speaking skills than in lessons on reading or grammar. Lessons on reading and grammar can be mediated through a tablet or other devices that display words or rules (thus combining the robot with other types of technology), while robots cannot fall back on such devices and need more skills (e.g., speech recognition, natural language generation) when practicing speaking skills with learners.

In contrast to the studies on language-learning outcomes, which showed mixed results, the studies addressing participants' learning-related emotions generally found positive effects, and showed that both children and adults often enjoy working with the robot. Given that learning motivation and learning gains are often related (Dörnyei, 1994; Pekrun et al., 2002), the robot's potential to motivate learners could be a valuable property. However, higher motivation was not always linked to higher learning gains in the RALL studies reviewed, and motivation could, at least in part, be due to the initial novelty effect of robots, which soon disappears. Although addressed in some experiments (e.g., Han et al., 2008), it is not clear how novelty has affected the results of previous studies. A strong recommendation is to carefully consider how to introduce the robot to participants to minimize novelty effects and to see whether the effects found are robust to prolonged exposure or wear off over time.

Conflicting results were especially striking with respect to the social behavior of the robot. Although some studies found positive effects of personalized and/or social behavior on learning gains and enjoyment, other studies found social behavior to negatively impact learning outcomes and behavior. Thus, it is clear that there is a thin line between the robot being social enough to sustain children's interest and being too social, leading to children being distracted or even intimidated by the robot. Furthermore, adaptivity and feedback have remained understudied and should have a more central role in future studies, given the importance of adapting to the learner's level and providing helpful feedback in L2 education and education in general (Li, 2010; Vygotsky, 1978).

One of the issues that makes it difficult to compare findings across studies is teleoperation. In the section on word learning, we discussed how teleoperation allows the experimenter to control the robot in real-time, which may result in different interactions than when the robot is running autonomously. A teleoperated robot can respond in a contingent manner, while autonomous robots have to rely on predefined scripts and can only respond contingently when the participant behaves as expected. One of the values of RALL research is its contribution to developing autonomous robots that can be placed in classrooms. This does not mean, however, that robots should not be teleoperated during experiments, as teleoperating robots allows researchers to study more advanced interactions than those that can be achieved using autonomous robots. Such interactions are valuable to identify robot behaviors or properties that need to be developed further. However, we recommend that researchers clearly state whether they teleoperated their robot or used it autonomously to make the distinction between the two types of robots more clear and to facilitate comparisons between studies.

A further important issue follows from the newness of the field: Robots constitute a new form of technology, and too few studies have been conducted so far to conclude that robots are effective language tutors. Future studies will allow for firmer conclusions regarding robots' potential as language tutors. Furthermore, a subset of the previous studies is heavily underpowered and/or suffers from other methodological limitations (e.g., no control group), which warrants caution in interpreting and evaluating their results. These issues often come to the fore in studying the use of technology for language learning: New technologies are often met with great enthusiasm, but research on their effectiveness often does not meet empirical standards and/or does not necessarily provide conclusive evidence of the benefits of these technologies (Salaberry, 2001). However, even with their limitations, the studies reviewed in this chapter helped us to identify potential areas for future research.

In our Introduction, several advantages of robots over many other forms of technology were discussed. One advantage is that robots provide opportunities for the learner to interact with their real-life environment. They are physically present and make it possible to integrate physical exercises or objects into learning tasks. Thus far, motor activities with robots have rarely been incorporated in learning tasks due to their feasibility (e.g., walking is undesirable as robots are likely to fall over), with the exception of a few studies (e.g., Tanaka & Matsuzoe, 2012). This is not surprising, because the more a robot acts and moves through space, the more likely it is that technical issues such as falling over or overheating will arise. However, it is possible that the use of objects and exercises could lead to higher learning gains (Kersten & Smith, 2002; Mavilidi et al., 2015; Toumpaniari et al., 2015). As robot technology is developing, motor activities are becoming more feasible to integrate, and this is therefore an area worth exploring.

Another advantage of robots over most other technologies is the interactional possibilities robots provide. Given the mixed evidence on the robot's social behavior, personalization of child–robot interaction is perhaps the most important line of research for the future. When a human teaches a child new word forms and meanings, they carefully monitor the child's comprehension and, if necessary, adapt the tutoring strategy to the child's needs. Robots are not yet capable of monitoring the child's comprehension and adapting the tutoring strategy to individual children in such a careful manner. This makes it difficult to obtain 'true' adaptivity, in which the robot adapts its lesson and behavior depending on the child's comprehension. This is partially due to speech recognition systems, which, in their current form, are often not suited to recognizing child speech (Kennedy et al., 2017), and to systems not being advanced enough to recognize children’s emotions or comprehension. Furthermore, it is difficult to program robots to respond in a contingent manner, and even more so in interactions with young children, who are much more unpredictable in their verbal responses than older speakers. In other words, the research so far indicates that the important advantages of robots over most forms of technology still exist primarily in theory. Technical limitations prevent regular implementation of these possible advantages, and further technological developments are required to make full use of robots’ potential and to put these theoretical advantages into practice.

A review of research on computer-assisted language learning (Garrett, 2009) mentioned how, in 1991, it was possible for one person to write "an overview... of the kinds of technological resources currently available to support language learning and of various approaches to making use of them" (Garrett, 1991, p. 74, as quoted in Garrett, 2009). In 2009, the same author noted that an update would fill an entire journal, requiring contributions from many different areas of expertise (Garrett, 2009). Perhaps in another two decades, we will say the same about RALL. We will go from one review aimed at capturing almost all extant RALL research to a great many possibilities we cannot even imagine at the moment. The use of robots may become such an everyday aspect of life that we will not even wonder about employing them. However, before we reach that point, we first have to find out how exactly robots should interact and behave socially to be effective language tutors.

Appendix

Table A.1. Robots Used in the Various Studies and Their Characteristics

Robot	Research or company based	Developed by	Target age group	Purpose	Body	Face	Arms
Bioloid	Company	Robotis	All ages	Build your own robot	A.o. humanoid, spider, or dinosaur	Fixed	Yes, movable by robot
Dragonbot	Research	MIT Media Lab	Children	Long-term interactions with children	Shaped like a dragon	Animated on a screen	Yes, to be posed by others
EngKey	Company-made for research purposes	Korea Institute of Science and Technology	Children	Real-time interaction between students and their remote teachers	Semi-humanoid	Avatar on a screen	Yes, movable by robot
iCat	Company-made for research purposes	Philips Electronics	All ages	Research on human-robot collaboration, joint attention, gaming, and ambient intelligence	Shaped like a cat	Movable eyes, eyebrows, lips	No
iROBi(Q)	Company	Yujin Robot	Children	Early childhood education	Semi-humanoid	Fixed, with emotion expression through colors	Yes, movable by robot
MecWilly	Company	MecWilly Project	All ages	Education	Humanoid	Movable eyes, eyebrows, lips	Yes, movable by robot
Mero	Company-made for research purposes	Korea Institute of Science and Technology	All ages	Various types of interactions	No	Movable eyes, eyebrows, lips	No
Nabazag	Company	Violet/Mindscape/ Softbank Robotics	All	Smart object	Shaped like a rabbit	Fixed, with movable ears	No
NAO	Company	Softbank Robotics	All ages	Various types of interactions	Humanoid	Fixed	Yes, movable by robot
Robo Sapien	Company	WowWeeBio-loidRoo	All ages	Making programming robots easy	Humanoid	Fixed	Yes, movable by robot
ROBOSEM	Company	Yujin Robot	All ages	L2 English teaching	Humanoid	An avatar on a screen	Yes, movable by robot
Robovie (R3)	Company-made for research purposes	Vstone & ATR	All ages	Assistance in everyday tasks	Humanoid	Fixed	Yes, movable by robot
Tangible robot companions	Research	Wang et al. (2013)	Children	L2 English teaching	Animals	Fixed	Yes, movable by robot

3

Tablets versus toys: Investigating the effect of using virtual or physical objects on children's L2 word learning

Rianne van den Berghe, Josje Verhagen, Ora Oudgenoeg-Paz, Sanne van der Ven, Fotios Papadopoulos, & Paul Leseman. (in preparation).

Author contributions: R.B., J.V., O.O., S.V., and P.L. conceptualized the research; F.P. developed the tablet environment; R.B. conducted and supervised the data collection, analyzed the data, and wrote the paper; J.V., O.O., S.V., and P.L. supervised; J.V., O.O., S.V., F.P., and P.L. critically reviewed the paper.

Abstract

In line with an embodied cognition view, one would expect that children learn more from interacting with physical objects rather than from interacting with virtual objects on touch screens. The present experiment compares the effect on children's L2 word learning of manipulating either physical objects or 3D models of these objects on a tablet screen during a vocabulary training. Ninety-nine Dutch preschoolers participated in an L2 vocabulary training that was aimed at teaching six English words and, crucially, either allowed children to manipulate physical objects or virtual objects to act out these target words. Word learning was assessed directly after the training and one week later via two word-knowledge tasks. Contrary to our expectations, we found that children presented with virtual objects performed equally well as children presented with physical objects on both word-knowledge tasks during both the immediate and the delayed post-test. Irrespective of condition, children obtained higher scores in the delayed post-test than in the immediate post-test on the word-knowledge tasks. Taken together, these results indicate that manipulating virtual objects on tablets do not affect preschool children's L2 word learning differently than physical objects, and that children need time to consolidate word knowledge.

Keywords: tablet, physical objects, embodiment, second language vocabulary, preschool children

Overview

As technologies are developing, they are increasingly incorporated into second-language (L2) education. L2 teachers make use of tablets, interactive white boards, videos, computer games, and many other forms of technology (Golonka, Bowles, Frank, Richardson, & Freynik, 2014). While some forms of technology (e.g., pronunciation training through automated speech recognition systems) look promising, the effectiveness of most forms of technology has not been established yet (Burnett, 2010; Golonka et al., 2014). Professionals express the concern that especially for young children, the use of technology is not natural or developmentally appropriate (Miller, 2005). As evidence is scarce, research is needed on the effects of specific technologies for educational purposes for young children (Burnett, 2010).

The focus of this article lies on the use of physical versus virtual objects displayed on a tablet screen to teach young children L2 words. Tablets can potentially be used to teach young children L2 words: Children can play interactive games on tablets, and thus, become familiar with L2 words in a playful and interactive manner. However, there could be a possible drawback to the use of tablets to teach young children L2 words – the fact that children work within a virtual environment rather than a physical one. For first language (L1) learning, interactions with the physical environments are thought to stimulate children’s language development through sensorimotor interactions (Barsalou, 2008; Hockema & Smith, 2009; Iverson, 2010; Wellsby & Pexman, 2014). For example, physically sorting objects likely contributes to learning semantic categories (Hockema & Smith, 2009; McGonigle & Chalmers, 2001) and research suggests that performing the relevant action while learning a novel verb is beneficial for learning (James & Bose, 2014; James & Swain, 2011).

Thus, previous work suggests that L1 learning is positively impacted by the use of physical objects as learning materials. Physical objects allow for more elaborate sensorimotor interactions as compared to virtual objects. While manipulating virtual objects, children cannot, for example, feel the weight or texture of an object. However, it is an open question whether this is also the case for L2 learning. To date, there has been no research on this topic yet. To fill this gap, this study compares the effect on children's L2 word learning of manipulating physical objects versus manipulating 3D models of such objects on a tablet screen. In so doing, the study builds on previous literature within the embodied cognition approach, discussed below.

The Embodied Cognition Approach

According to the embodied cognition approach, sensorimotor interactions, such as reaching, grasping, manipulating, sorting, inserting, stacking, and changing spatial relations are important to learn, understand, and use language (Barsalou, 2008; Hockema & Smith, 2009; Iverson, 2010; Wellsby & Pexman, 2014). Through these basic interactions children learn about their environment and link words to referents. In this view, vocabulary knowledge is grounded in basic sensorimotor interactions.

Hockema and Smith (2009) view language learning as a combination of outside-in and inside-out processes. On the one hand, children learn by perceiving the linguistic and physical regularities in their environment. For example, when round objects are constantly called balls, children learn this word due to the linguistic regularity (same word) and physical regularity (physically similar referents). On the other hand, children also play an active role in this process, as they act on their environment and produce such regularities, for example by grouping balls together while playing, grasping objects such as balls to explore their physical attributes, and moving around to explore spatial relations (Gibson & Pick, 2000; Hockema & Smith, 2009; Oudgenoeg-Paz, Volman, & Leseman, 2016; see also Gogate and Hollich, 2010, for a discussion of how children actively elicit linguistic regularities that change with development). Thus, children make use of statistical learning to learn words and word meanings, by keeping track of statistical (semantic and physical) regularities to pair words to referents (e.g., Smith & Yu, 2008). At the same time, children also play an active role in generating these regularities (Gogate & Hollich, 2010).

An important tenet within the embodied view of language development is that children construct concepts of objects and their affordances through their sensorimotor interactions with the environment (Antonucci & Alt, 2011). Affordances are possibilities for action and as such are defined by both the properties of the object or environment and the child's skills (Gibson, 1986). Children see, touch, and move objects, and, therefore, the concepts they construct include these sensorimotor experiences of objects' features and affordances. Word learning involves mapping labels onto concepts, that, in turn, are grounded in real-life sensorimotor interactions (Gallese & Lakoff, 2005; Howell, Jankowicz, & Becker, 2005; Scofield, Hernandez-Reif, & Keith, 2009). Therefore, sensorimotor experiences are part of the lexicon (Öttl, Dudschig, & Kaup, 2017). Importantly, sensorimotor experiences do not just play a role in initial word learning, but remain part of the lexicon and are unconsciously and automatically retrieved during language processing (Myung, Blumstein, & Sedivy, 2006).

Previous research taking an embodied cognition approach has mostly focused on L1 learning rather than L2 learning. Therefore, it is currently not clear whether sensorimotor interactions play an important role in L2 learning as well. The hypothesis to be tested in our study is that L2 learners benefit from sensorimotor interactions, just like L1 learners, because these interactions would help L2 learners in linking L2 labels to the underlying embodied concept. Specifically, sensorimotor interactions would activate the embodied concept that the learner has already acquired in their L1, and would make it easier to link the L2 word to this concept. Indeed, initial evidence for this idea comes from work showing that the use of iconic movements or gestures in L2 word learning leads to higher learning gains (Mavilidi, Okely, Chandler, Cliff, & Paas, 2015; Tellier, 2008; Toumpaniari, Loyens, Mavilidi, & Paas, 2015). Specifically, these studies showed that preschool children making iconic movements or gestures learned L2 words better than when making random movements or no movements at all. This use of movements or gestures is thought to help children activate the embodied concepts they have already acquired in L1 and thus assist in mapping the L2 word onto that concept.

However, potential benefits of other types of sensorimotor interactions, such as interacting with physical objects, has, to the best our knowledge, not yet been researched. One exception to this is an exploratory study, further discussed below, which compared the effects of using physical versus virtual objects on L1 word learning (Singer & Gerrits, 2015). In this study, three-year-old children were found to learn equally many words interacting with physical objects compared to virtual objects. In the absence of earlier work addressing children's L2 learning using either physical or virtual objects, we briefly discuss the main outcomes of a number of studies comparing children's use of physical and virtual objects, with a different focus: L1 learning, reading comprehension, and mathematics.

Language Learning from Virtual Environments

While it is clear that sensorimotor interactions help language learning, it is not clear yet whether children learn more from vocabulary trainings that include sensorimotor interactions as compared to trainings that do not, or to a lesser extent, allow for such interactions (e.g., when they include manipulations of virtual objects). Several studies indicate that children can learn the meaning of novel words through manipulating objects (e.g., O'Neill, Topolovec, & Stern-Cavalcante, 2002; Smith, 2005), but direct comparisons between physical and virtual objects have rarely been made. Rather, previous work has compared virtual tools to paper games to assess their effectiveness. For example, Lan, Fang, Legault, and Li (2015) investigated virtual environments for language learning, and compared adults' learning via either a virtual environment or via printed-out pictures in an L2 (Chinese) word learning training consisting of seven sessions. For the first three sessions, participants in the virtual environment condition were outperformed by the participants in the picture condition when assessed on a picture-selection task immediately after the training. However, the participants in the two conditions showed similar learning gains from the fourth session onwards and during a delayed post-test administered three weeks after the training. Similar results were obtained by Heitink, Fisser, and Voogt (2010) who found that nine- to twelve-year-old children playing an online game aimed at improving their L1 vocabulary outperformed children playing a paper game, but only when combined with classroom vocabulary instruction. The advantage of an online game when combined with classroom vocabulary instruction was found during both an immediate post-test and during a delayed post-test, four weeks after the training.

It is not clear how the findings of Lan et al. (2015) and Heitink et al. (2010) relate to learning of new L2 labels for L1 forms by using physical objects. Both studies used pictures rather than physical objects. Clearly, pictures do not allow for the same kind of sensorimotor interactions as physical objects. To the best of our knowledge, only one (unpublished) study to date has compared the effects of using physical versus virtual objects on word learning. In this study by Singer and Gerrits (2015), which looked at L1 word learning, three-year-old children were taught five words through a tablet game in one session, and five words through playing with physical objects in another session. Children learned one new word on average during each of the two sessions, as measured via a picture-selection task one week later. Children did not show any learning on a picture-selection task during a post-test immediately after the training. Crucially, moreover, the two conditions did not affect children's learning differently. These findings are unexpected from the viewpoint of embodied cognition

(Hockema & Smith, 2009; Wellsby & Pexman, 2014), but are in line with other types of learning for which direct comparisons have been made between physical and virtual objects: reading comprehension and mathematics.

Comparing Effects of Physical and Virtual Objects on Reading Comprehension and Mathematics

Previous research has shown that the possibility to interact with physical materials in regular classes, which often do not incorporate the possibility of such interactions, enhances learners' understanding of the subject for both mathematics and reading comprehension (Glenberg, Brown, & Levin, 2007; Uttal, Scudder, & DeLoache, 1997). However, when studies directly compared the use of physical and virtual objects, no differences between conditions were found, similar to Singer and Gerrits (2015). For example, Glenberg, Goldberg, and Zhu (2011) found that manipulations on a computer were just as beneficial for six- to eight-year-old children's reading comprehension as the use of physical objects. In fact, even imagined manipulations were found to be as effective as manipulations of physical objects for reading comprehension and mathematics (Glenberg, Gutierrez, Levin, Japuntich, & Kaschak, 2004; Glenberg, Willford, Gibson, Goldberg, & Zhu, 2011). These findings suggest that manipulations of any kind (thus, with physical objects, virtual objects, or even imaginary) provide children with enough opportunities to increase learning or comprehension, at least when aimed at increasing children's math or reading skills. However, reading comprehension and mathematics are very different from vocabulary in many respects. Retrieving embodied concepts aids in reading and doing mathematics, while the learners has to actively put a new label on such concepts when learning (L2) vocabulary. It is not clear whether findings from reading comprehension and mathematics extend to (L2) word learning.

This Study

The present study investigated whether children's manipulation of physical objects versus virtual objects on a tablet screen affects their L2 word learning differently. Even though technology is widely used in L2 education, its effects are still unclear (Burnett, 2010; Golonka et al., 2014). Hence, it is crucial to investigate whether physical objects have a benefit over virtual objects, and whether virtual objects do not impact L2 word learning negatively. In our study, children were presented with a vocabulary training in which they used either a tablet or physical objects to experience the meanings expressed by the target words. Children's word-learning gains were measured directly after the training and one week later. The delayed post-test assessed children's retention of the target words, and was included because studies show that learned vocabulary needs time to consolidate, resulting in higher learning gains measured after a short delay than immediately after a training (for an overview, see Axelsson, Williams, & Horst, 2016). Sleep plays an important role in the consolidation of new words, as sleep is an active process that helps strengthen newly acquired information (Axelsson et al., 2016; Diekelmann, Wilhelm, & Born, 2009; Stickgold & Walker, 2013). In fact, we wanted to explore the possibility that the beneficial effect of using physical objects over virtual objects was more pronounced one week after the training. We controlled for phonological memory during our analyses, using a nonword repetition task, as nonword repetition ability is known to be related to language learning ability, in particular word learning, in both in L1 and L2 (for an overview, see Gathercole, 2006).

We expected that children in the object condition would outperform children in the tablet condition on the word-knowledge tasks during both the immediate post-test and the delayed post-test, as the sensorimotor interactions of the children with the objects provided by the object condition would result in greater learning gains compared to the tablet condition. For example, one of the L2 target words was 'heavy', and we expected an advantage of holding a heavy object over swiping a model on a tablet screen when learning this word. The sensorimotor interactions with the heavy object could activate the child's embodied concept of 'heavy', making it easier for the child to connect the L2 word to the concept of 'heavy'. Swiping a model on the tablet screen would not lead to the same activation of the concept 'heavy', thus making it more difficult to connect the L2 word to its concept. Furthermore, we expected children to obtain higher scores on the word-knowledge tasks in the delayed post-test than in the immediate post-test, in line with research on consolidation in word learning (Axelsson et al., 2016).

Method

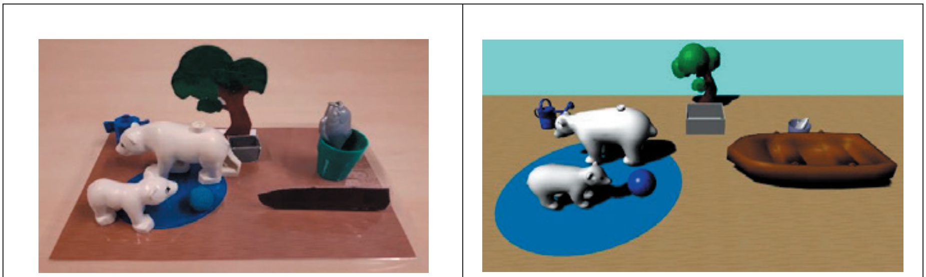
Participants

Ninety-nine preschoolers (50 girls) with an average age of 64.7 months (range 48.7 – 77.3 months, *SD* = 7.5 months) participated in the study. They were tested at various kindergartens in the Netherlands. Within schools, half of the children (*n* = 52) were randomly assigned to the tablet condition (*M* = 64.6 months, age range = 50.3 – 77.3 months, *SD* = 7.4 months; 27 girls) and the other half (*n* = 47) to the object condition (*M* = 64.9 months, age range = 48.7 – 75.1 months, *SD* = 7.6 months; 23 girls). Of these 99 children, three children in the object condition were excluded from analysis as they already knew one or two of the target words, as measured during the pre-test discussed below. One of the children in the tablet condition completed the vocabulary training but could not participate in the immediate post-test due to extracurricular activities. All children participated in the delayed post-test. Informed consent for all children was obtained from parents/caregivers prior to data collection. All parents filled out a questionnaire regarding children's language use and development, parental education, motor development, and experience with using tablets. Most children had experience with touch screens and all children were able to perform the basic manipulations with the physical or virtual objects, such as lifting the physical objects and swiping the virtual objects on the tablet screen, as indicated by their parents. Seventeen children had another home language next to Dutch. Two children, both in the object condition, had English as one of their home languages and were excluded from analysis. For the other children in the study, parents reported that they did not have any knowledge of English. This was supported by the pre-test data, which indicated that the children did not know any of the target words.

Vocabulary training

During the vocabulary training, children were presented with six English words: 'heavy', 'light', 'full', 'empty', 'in front of', and 'behind'. Contrary to Singer and Gerrits (2015), we selected target words such that their meaning was clearly grounded in physical experiences, to maximize the potential benefits for physical objects over virtual objects. The words 'heavy' and 'light' are grounded in children's experiences with weight, through holding heavy and light objects. The words 'full' and 'empty' are grounded in children's experiences with filling and emptying containers, such as cups. Last, the words 'in front of' and 'behind' are grounded in children's experiences with spatial relations. While the concepts of 'in front of' and 'behind' can be explored from different viewpoints with physical objects, this is more difficult for virtual objects.

The six target words were embedded in a narrative task, in which the child was instructed that they had to take care of animals in a zoo. The children were then told a story by a trained assessor in which each target word was presented ten times, in varying contexts. Specifically, the child was asked to repeat the target word upon first exposure, and had to answer a Dutch to English translation question during the training (e.g., "what was *zwaar* [heavy] again in English?") in both conditions. Children were encouraged to repeat the target words, and were provided with the translation of the target word if they were not able to answer the translation question. See Figure 1 for a fragment of the training, and the Appendix for the whole narrative.



(The child is asked to put the polar bears in front of the boat.)

The polar bears are now in front of the boat. The English word for in front of “**in front of**”. {If the child does not repeat the word, say: “Repeat after me: **“in front of”**}. Good job, **in front of!** So what is “in front of” in English? {If the child does not say the target word, say: In front of is in English “**in front of**”}. There is a bucket with fish “**behind**” the boat. **Behind** means “behind”. {If the child does not repeat the word, say: “Repeat after me: “**behind**”}. Good job, **behind!**

Figure 1. Example of the physical-object condition (left) and the virtual-object condition (right), with corresponding excerpt from the training.

Crucially, manipulations of the objects representing the attribute (e.g., a Duplo elephant filled with sand to make the toy relatively heavy) were required at various fixed moments during the training. These manipulations were kept exactly the same across conditions, except that in the object condition, children performed the actions with physical materials (e.g., placed the heavy Duplo elephant in a toy cage), while, in the tablet condition, children manipulated 3D models of these same objects on the screen (e.g., swiped a similarly sized 3D model of the same Duplo elephant into a cage). At the end of the narrative, children were presented with a summary of the story in which the six target words were repeated once more (for the tenth time), to minimize differences in time elapsed since hearing the target words last and the test phase. The training lasted 15-20 minutes in total.

Measures

Translation task

A translation task was used in the pre-test and post-tests. In this task, children were asked to translate the six target words from English to Dutch (e.g., "What is *light* in Dutch?"), and vice versa (e.g., "What is *licht* [light] in English?"). The aim of this task was to assess children's knowledge of the word forms, that is, whether they could produce the L1 or L2 counterpart of a word. Full points were awarded for correct answers, yielding a maximum score of twelve. As a pre-test, each child was asked to translate the words from English to Dutch, to assess whether the child already knew any of the target words prior to the training. Cronbach's alpha showed that the reliability of the translation task was acceptable for the immediate, α = .71, and delayed post-test, α = .67.

Comprehension task

To measure children's receptive knowledge of the target words, a picture-selection task was used in which children were asked to select one picture (out of four options) which best matched a target word. There were four items per target word, yielding a total of 24 items. For each item, children responded to the question "Where do you see X?", asked by the assessor. For the two prepositions (i.e., 'behind', 'in front of'), the question was "Where do you see X [e.g., the dog] behind/in front of Y [e.g., the house]?", in order to make clear which aspect was the focus of that picture and was to be compared to its context. The questions were asked in Dutch (L1) and only the target words were in English (L2). Children were allotted one point for each correct answer, yielding a maximum score of 24. The comprehension task was administered to a subgroup of 50 children. This was because the first group of children scored below chance level on a few items, suggesting that the target picture was not clear enough or that there were issues with distractor items. Therefore, an adapted version of the task was used for the remaining 50 children. Cronbach's alpha showed that the reliability of this task was just below acceptable for both the immediate, α = .66, and delayed post-test, α = .65.

Non-word repetition task

As a control measure, a non-word repetition task was used to measure phonological memory. This task contained a sub-set of the items used in Rispens and Baker (2012), which were embedded in a computerized task appropriate for young children (for more details, see Verhagen, de Bree, Mulder, & Leseman, 2017). In this task, children were presented with a previously recorded, non-existing word via a laptop computer, that they were asked to repeat. The task contained twelve items. Before starting this task, there were two practice items (one Dutch word and a two-syllable nonword), for children to practice repeating the items. Children were scored online by the assessor and rewarded one point for each correctly repeated word, thus the maximum score on this task was twelve. Cronbach's alpha showed that the reliability of this task was just below acceptable, α = .61, but α = .83 in the original paper (Verhagen, de Bree, et al., 2017). Ten percent of the data (ten videos) was scored by an additional researcher. Inter-rater reliability was good with 82% agreement, κ = .73 (95% CI, .611 to .855), $p <$ .001. Consensus was reached on items that had been scored differently.

Additional measures

Two additional measures used but not included in the current chapter were a sorting task and a story recall task. The sorting task was used as an additional comprehension task. However, this measure was not included in the analyses, as the test-retest reliability of this measure was inadequate (*r*(49) = .041, *p* = .779). A story recall task was used to measure children's recall of the narrative. This measure is beyond the scope of the current chapter as it did not measure children's word learning.

Procedure

All children were tested individually in a quiet room in their schools by a trained assessor. The first session lasted about 40 minutes, and the second session about 30 minutes. In the first session, the pre-test (the English-to-Dutch translation task) was administered first to assess whether children already knew the target words. Following the pre-test, children were allowed to freely play with the tablet or objects (depending on the condition) before starting the training, to allow them to get used to them and therefore, make sure they would be focused on the contents of the training instead of the materials themselves. This also allowed the children in the tablet condition to practice operating the tablet. When children had finished playing (which never took more than five minutes), the training started. After the training, the tasks of the immediate post-test were administered, in the following order: comprehension task, (sorting task,) English-to-Dutch translation task, (story recall task,) and Dutch-to-English translation task. One week later, the delayed post-test session was conducted to measure children's retention of the target words. In this session, the five tasks were administered again, in the same order as in the immediate post-test, followed by the non-word repetition task. In both sessions, children got a sticker as a reward for each task completed. At the end of each session, they received a small gift.

Analyses

We ran independent-samples t-tests to compare the two groups of children on age and phonological memory, as well as a Pearson's Chi-Square Tests to investigate any differences in gender composition between the two groups. Children's scores on the comprehension task were compared against chance level, which was 25%.

To investigate differences in learning gains between the children learning with virtual objects and children learning with physical objects, we ran mixed-effect logistic regression models in the statistical package R (R Core team, 2017) and the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). Mixed-effects models can deal both with within- and between-subject variables (e.g., post-test and condition), and with variance across and within learners (e.g., different trajectories from the immediate to the delayed post-test). In these models, scores on the translation task and comprehension task were the dependent variables. 'Subjects' and 'items' were included as random factors, and random slopes for subjects (post-test) and items (condition*post-test) were included. The factors 'condition' and 'post-test' were included as fixed effect factors. For the translation tasks, 'language' (from English to Dutch or from Dutch to English) was included as an additional factor. We did not include bilingualism as a factor in our main analyses. Rerunning our analyses with bilingualism as an additional factor did not yield different results from our main analyses without bilingualism, and bilingualism did not affect children's scores on the word-learning tasks.

Results

Descriptive Analyses

Independent-samples t-tests indicated that the children in the two groups were comparable in age, $t(91) = .287$, $p = .774$, and phonological memory, $t(91) = -.110$, p

= .913. A Pearson Chi-Square Test indicated no significant differences in gender composition between the two groups, $\chi^2(1, N = 94) = .361$, $p = .548$. For the subgroup of 50 children (*M* = 68.3 months, age range = 48.7 – 77.3 months, *SD* = 6.3 months; 23 girls) who had taken the comprehension task, independent-samples t-tests revealed that – unlike for the larger sample – children in the two groups differed significantly in age, $t(43) = 2.424$, $p = .020$, $d = .67$, such that the children in the tablet condition (*M* = 70.5 months, *SD* = 4.2 months) were older than the children in the object condition (*M* = 66.5 months, *SD* = 7.1 months). The children did not differ in phonological memory, $t(48) = .363$, $p = .719$, or gender $\chi^2(1, N = 50) = .057$, $p = .811$. Hence, in our analyses of the comprehension task for this specific sub-sample, age was included as an additional fixed effect.

Table 1 displays the mean task scores on the translation task and comprehension task for both the immediate and the delayed post-test for the two conditions separately. One-sample t-tests were conducted to compare children's mean scores on the comprehension task against chance level. Children in the object condition scored significantly above chance level on the comprehension task during both the immediate post-test, $t(26) = 10.303$, $p < .001$, $d = 1.98$, and delayed post-test, $t(26) = 11.238$, $p < .001$, $d = 2.16$. Children in the tablet condition also scored

Table 1. Mean Task Scores (SD) on the Translation Tasks and Comprehension Task

Session	Condition	English-Dutch	Dutch-English	Comprehension task
Immediate	Physical	1.38 (1.21)	0.90 (1.03)	12.44 (3.25)
	Virtual	0.98 (1.12)	0.65 (1.00)	12.27 (3.78)
Delayed	Physical	1.83 (1.41)	1.45 (1.13)	14.26 (3.82)
	Virtual	1.73 (1.43)	1.19 (1.14)	14.30 (3.44)

Note. The maximum score was six for each translation task, and 24 for the comprehension task.

significantly above chance level on the comprehension task during both the immediate post-test, $t(21) = 7.780, p < .001, d = 1.66$, and the delayed post-test, $t(22) = 11.566, p < .001, d = 2.41$.

Table 2 displays the correlations between the tasks for both post-tests. The moderate correlations between the translation tasks and comprehension task indicate that the tasks measure related, yet distinct types of vocabulary knowledge. The correlations between the two post-tests for each measure are moderate to strong, indicating adequate test-retest reliability.

Table 2. Correlations between the Scores on the Various Word-Knowledge Tasks

Session	1	2	3	4	5	6
1. English-Dutch (immediate)						
2. Dutch-English (immediate)	.617***					
3. Comprehension (immediate)	.517***	.428***				
4. English-Dutch (delayed)	.697***	.429***	.500***			
5. Dutch-English (delayed)	.627***	.552***	.348**	.638***		
6. Comprehension (delayed)	.464***	.478***	.614***	.600***	.463***	

Note. ** $p < .01$; *** $p < .001$.

Word-Learning Tasks

First, we investigated the effect of physical and virtual objects on children's scores on the translation task. A model with 'subjects' and 'items' as random effects, random slopes for subjects (post-test) and items (condition*post-test), and language, condition*post-test as fixed effects showed a main effect of post-test ($ß = 1.01$, $SE = .21$, $z = 4.71$, $p < .001$), and language ($ß = -.69$, $SE = .13$, $z = -5.44$, $p < .001$), but no main effect of condition ($ß = -.51$, $SE = .35$, $z = -1.44$, $p = .149$), or interaction effect between post-test and condition ($ß = .51$, $SE = .32$, $z = 1.58$, $p = .114$). These effects indicated that children obtained higher scores during the delayed post-test than during the immediate post-test, and that children obtained higher scores on the English-to-Dutch questions than on the Dutch-to-English questions, irrespective of condition.

Next, we compared children's scores on the comprehension task, which had been administered to a subgroup of 50 children (*n* tablet condition = 23, *n* object condition = 27). A model with 'subjects' and 'items' as random effects, random slopes for subjects (post-test) and items (condition*post-test), and age, condition*post-test as fixed effects showed a main effect of post-test (*ß* = .41, *SE* = .13, *z* = 3.09, *p* = .002), but no main effect of condition (*ß* = -.06, *SE* = .24, *z* = -.24, *p* = .814), age (*ß* = .02, *SE* = .02, *z* = 1.46, *p* = .144) or interaction effect between post-test and condition (*ß* = .05, *SE* = .19, *z* = .25, *p* = .801). Irrespective of condition, children obtained higher scores during the delayed post-test than during the immediate post-test.

We re-ran the models for scores on the translation task of the subgroup of 50 children to check whether the findings of the full sample also held for the smaller sample. A model with 'subjects' and 'items' as random effects, random slopes for subjects (post-test) and items (condition+post-test), and language, age, condition*post-test as fixed effects yielded the same results as the model for the full group. The model showed a main effect of post-test (*ß* = 1.67, *SE* = .50, *z* = 3.33, *p* < .001), and language (*ß* = -.86, *SE* = .22, *z* = -3.97, *p* < .001), but no main effect of condition (*ß* = .13, *SE* = .59, *z* = .21, *p* = .83), age (*ß* = .06, *SE* = .03, *z* = 1.66, *p* = .097), or interaction effect between post-test and condition (*ß* = .06, *SE* = .54, *z* = .12, *p* = .909).

In summary, no effects of condition were found in any of the models, meaning that there were no significant differences in performance between the two conditions. An effect of post-test was found across vocabulary measures: Children performed better during the delayed post-test than during the immediate post-test. Also, children performed better on the English-to-Dutch translation task than on the Dutch-to-English translation task.

Discussion

The aim of our study was to investigate whether Dutch preschoolers who had no prior knowledge of English would learn L2 English words better when manipulating physical objects (that referred to the target words) during an L2 vocabulary training than when manipulating 3D models of these objects on a tablet screen. Following the embodied cognition approach (Barsalou, 2008; Hockema & Smith, 2009; Iverson, 2010; Öttl et al., 2017; Wellsby & Pexman, 2014), we expected that children learning with physical objects would outperform children learning with virtual objects. Specifically, we predicted that children's sensorimotor interactions with, for example, heavy objects when learning the word 'heavy' would result in higher learning gains as compared to interactions with virtual objects which do not allow children to experience a notion such as 'weight'. Contrary to our expectations, children who manipulated physical objects within a vocabulary training did not outperform children who manipulated 3D virtual objects on a tablet screen on any of the word-learning tasks during both the immediate and the delayed post-test.

The lack of differences in word knowledge between the two groups of children suggests that children's manipulations of physical objects or virtual objects on a tablet screen do not affect L2 vocabulary learning gains differently. Above-chance performance on our tasks indicated that children were able to learn words from the vocabulary training, even though learning rates were rather low overall. Our findings contrast with research on the importance of sensorimotor interactions for language learning. In the embodied cognition approach, sensorimotor interactions are thought to be important to learn, understand, and use language, as concepts are grounded in one's interactions with the environment (Hockema & Smith, 2009; Öttl et al., 2017; Wellsby & Pexman, 2014). One would expect that word learning situations would benefit from interactions that help activate underlying concepts and connect L2 words to these concepts. For example, children in our study could have benefited from lifting a heavy Duplo elephant while learning the word 'heavy', compared to children swiping a virtual elephant on a tablet screen. The heavy physical elephant should have helped activate the underlying embodied concept of 'heavy' more than the virtual elephant. Previous work has shown that actions that pertain to the target word, such as, iconic movements or gestures performed by the learner benefit L2 word learning more than random movements or no movements at all (Mavilidi et al., 2015; Jalongo & Sobolak, 2011; Toumpaniari et al., 2015). Manipulation of physical objects was expected to have a similar effect on L2 word learning. However, no such benefit was found in our study.

A possible explanation for our results is that manipulations of any kind, either of physical or virtual objects, benefit children's learning. Our findings are similar to those of research on the use of physical and virtual objects in mathematics and reading comprehension (Glenberg, Goldberg, et al., 2011; Glenberg et al., 2004; Glenberg, Willford, et al., 2011). Manipulations on a computer, and even imagined manipulations, were found to be just as beneficial for reading comprehension and mathematics as the use of physical objects. Glenberg et al. (2004) explain the benefits of computer or imagined manipulations on reading comprehension in terms of the Indexical Hypothesis, which states that words become meaningful through sensorimotor simulation of the content of sentences (Glenberg, 1997). Computer and imagined manipulations help the reader to simulate sentences, and therefore, understand them better. This differs from word learning, as children in our study needed to retrieve one specific concept rather than combining several concepts of one sentence. Specific to our study, the manipulations on the tablet screen may have helped children to simulate movement, and thus, simulate concepts such as weight. These simulations may have activated the underlying concept of the word, and thus, helped children to learn the novel L2 word. Taken together with previous research, our study seems to indicate that virtual objects, at least in their 3D form, as used in our study, can provide children with enough opportunities for learning L2 vocabulary.

Another possible explanation for the lack of differences is that we studied L2 word learning as opposed to L1 learning. L2 word learning may be less dependent on sensorimotor interactions with objects, as learners have already acquired the concept in their L1. We know from studies about L1 learning that children first construct concepts, and then proceed to map labels onto these concepts (Antonucci & Alt, 2011; Howell et al., 2005; Scofield et al., 2009). When learning L1 words, both a word form and a concept have to be learned, while in our study, learning the target words required mapping a new L2 form to an existing concept and L1 word form. For example, if the learner does not possess the concept 'heavy' yet, experiencing the concept 'heavy' by holding something heavy may be crucial to acquire the concept and the word. In our experiment, the children already had an embodied concept of 'heavy', acquired when they were learning their L1. Thus, sensorimotor interactions may be more important to build conceptual representations than when mapping labels onto representations. In other words, it will play a more important role when learning an L1 than when learning an L2. Future research could look into L1 word learning using objects or tablets, or L2 words of which the concepts do not match the L1 concept the child has acquired. Learning L1 words or L2 words that do not overlap existing concepts may benefit more from interactions with physical objects than virtual objects on a tablet screen, and it remains an open question whether the lack of effects holds for such words.

In addition to the findings discussed above, we found that, on both word knowledge tasks, children performed better during the delayed post-test administered one week after the vocabulary training than during the post-test administered immediately after the training. These findings are in line with earlier results on word learning (see Axelsson et al., 2016 for a review). The novel words need to be consolidated: They need to be integrated into children's existing memory and children's knowledge of these new words need to be strengthened. Sleep plays an important role in the consolidation of new words, as sleep is an active process that helps strengthen and generalize newly acquired information (Diekelmann et al., 2009; Stickgold & Walker, 2013). A strong recommendation for future research on word learning, therefore, is to include a delayed post-test, which likely is a better measure of learning gains.

A strength of our study is that we tried to optimize the likelihood of finding any effects of using physical objects over virtual ones, in three ways. First, we included a delayed post-test, to explore the possibility that consolidation effects would show up differently across conditions. Second, by using various types of tests to measure word knowledge and administering these tests twice, we maximized the chance of finding effects, perhaps pertaining to specific types of knowledge. Finally, we selected words that were clearly grounded in embodied concepts such as 'heavy' and 'full'. Since we could not find benefits of physical objects for these words, it is unlikely that physical objects would benefit word learning for concepts that are more abstract and less clearly grounded. Limitations of our study are that we could only teach the children six words during our single vocabulary-training session, that the reliability of the receptive task was just below acceptable, and that the receptive task was administered only to a subgroup of our participants. Further research could study different types of words (e.g., verbs and nouns) to draw stronger conclusions on the use of virtual versus physical objects for L2 word learning. Moreover, possible benefits of physical over virtual objects are likely more pronounced in L1 word learning, which should be studied further. Our tasks did not measure the extent of children's semantic networks of the newly learned L2 words. Especially when comparing the effects of physical versus virtual objects on L1 word learning, researchers should consider measuring both how many words children learn and the extent of their semantic networks. Perhaps children do learn as many L1 words when playing with virtual objects compared to when playing with physical objects, but the semantic network of children using physical objects may be richer than that of children using virtual objects.

To summarize, we found no differences in L2 word learning gains for preschoolers manipulating physical objects during an L2 vocabulary training compared to preschoolers manipulating 3D models of these objects on a tablet screen. While we do not want to claim that tablets are just as effective learning tools as physical objects for all learners and all domains, it does seem to be the case that manipulating 3D objects on a tablet does not affect L2 word learning differently than physical objects.

Appendix

This appendix contains a translated version the narrative of the vocabulary training. The original narrative was in Dutch. Only the words in **bold** were said in English.

Narrative

Look, there is a zoo over here. One of the caretakers of the animals has fallen ill, so he cannot make it to the zoo today. You are asked to replace him. That means you will take care of the animals today. That is fun!

The zoo is in England. In England, people speak a different language from Dutch, namely: English. You will also learn some English words today. That is useful, because then you will know the language of the zoo a bit. Listen closely to the story and we will see later which English words you learned!

First, we are going to check on the elephants. Look, here are a mommy and a daddy elephant. The mommy elephant is the elephant with the pink bow. The elephants need to be put in their cage. The mommy elephant has a baby in her belly, so she is very heavy. The English word for “heavy” is “**heavy**” {If the child does not repeat the word, say: “Repeat after me: “**heavy”**}. Very good: **heavy**! Put the “**heavy**” mommy elephant carefully on the other side of the fence, inside the cage. What is heavy in English? {If the child does not say the target word, say: Heavy is in English “**heavy**”}. Well done! The daddy elephant does not have a baby in its belly of course. He is not very “**heavy**”. The daddy elephant is light. In English “light” is **“light”** {If the child does not repeat the word, say: “Repeat after me: “**light”}**. Well done: **light**! Put the “**light”** daddy elephant next to the “**heavy”** elephant in the cage, behind the fence. What is light in English? {If the child does not say the target word, say: Light is in English “**light**”}. Well done, now the “**heavy**” mommy elephant and the “**light**” daddy elephant are in their cage.

Can you check whether there is enough food in their tray? Oh it is completely empty. The English word for empty is “**empty**”. {If the child does not repeat the word, say: “Repeat after me: “**empty”**}. Good job: **empty**! The food tray is “**empty**”. What is empty in English? {If the child does not say the target word, say: Empty is in English “**empty**”}. You can put the sticks with leaves in the tray. Now the tray no longer is “**empty**”. Now it is full. In English that is “**full**”. {If the child does not repeat the word, say: “Repeat after me: “**full”}**. Yes, good job: **“full”**! What is full in English? {If the child does not say the target word, say: Full is in English “**full**”}. Well done!

Now we are going to the polar bears. There is a big mommy polar bear and a little one. Can you take the mommy polar bear out of the water and in front of the boat? The mommy polar bear is **heavy.** The little polar bear cannot stay in the water by itself. Can you put it next to its mommy? The baby polar bear is not “**heavy**”; it is “**light**”.

The polar bears are now in front of the boat. The English word for in front of “**in front of**”. {If the child does not repeat the word, say: “Repeat after me: **“in front of”**}. Good job, **in front of!** So what is “in front of” in English? {If the child does not say the target word, say: In front of is in English “**in front of**”}. There is a bucket with fish “**behind**” the boat. **Behind** means “behind”. {If the child does not repeat the word, say: “Repeat after me: “**behind**”}. Good job, **behind!** The bucket is now full, because there is a very big fish in it: it is **“full”.** You can give the fish to the polar bears. The bucket is now empty, it is “**empty**”. Put the **“empty”** bucket back “**behind**” the boat. What is “behind” in English? {If the child does not say the target word, say: Behind is in English “**behind**”}. You can also clean up the pool, because it was rather “**full**” with the polar bears and all the things. Put the ball into the box in front of the tree, so the box “**in front of”** the tree. Good job, now the box is “**full**”. The watering can is completely empty, so it is very light to lift, very “**light”**. Put the watering can “**in front of”** the tree as well. You did a really good job! The pool is empty now, completely “**empty**”.

Look, now there are giraffes! Look at that, some of the giraffes are playing hide and seek with us! Do you see where they are? There is one giraffe over there, behind those trees! It is “**behind**” the trees. It was well hidden “**behind**” the trees. There is also a giraffe sleeping in front of the bushes. It lies **in front of** the bushes. Can you put both giraffes “**in front of”** their cage, so in front of the cage? Well done! Now you can give them some food. The food tray is still in the corner. It is **behind** the bush, so behind the bush. There is nothing in it yet, it is completely “**empty**”. It is still “**light”**. Put it “**in front of**” the giraffes, then it is easy for them to reach it while eating. Can you fill the food tray with leaves? The leaves are also very “**light**”. Good job! Now the tray no longer is **“empty”,** it is completely **“full”.**

It is evening! You can bring the polar bears to their sleep pool. Put the “**heavy**” mommy polar bear behind the fence, so **“behind”** the fence, into the pool. Good job! Put the baby polar bear also “**behind**” the fence into the pool. The polar bears get to play for a little while before they go to sleep. Put the ball also in the pool. Put it “**in front of**” the mommy polar bear, so in front of the mommy polar bear. The ball is very light, very “**light”**, and otherwise the wind may blow it away! Well done! The pool is now “**full**”, so full, with two polar bears and a ball. The food tray is “**in front of**” the pool. Can you put the sticks with leafs also in the food tray? Then it is “**full**”, in case the polar bears want to eat something tonight. Good job! Now the cage is completely “**full”** and they can go to sleep.

Wow you did such a good job! I will tell the caretaker tomorrow everything that you did. You fed the “**heavy**” mommy elephant and the “**light**” daddy elephant. Their food tray was completely “**empty**” and you made sure it was “**full**” again. You also put the polar bears “**in front of**” the boat so that you could feed them a fish. Then the giraffes were playing hide and seek, one of them was “**behind**” the trees, and you fed them. And you also brought the polar bears to bed. Very good job!

4

Comparing children’s L2 word learning between a vocabulary training without a peer, with a child peer, or with a humanoid robot

Rianne van den Berghe, Sanne van der Ven, Josje Verhagen, Ora Oudgenoeg-Paz, Fotios Papadopoulos, & Paul Leseman. (in preparation).

Author contributions: R.B., J.V., O.O., S.V., and P.L. conceptualized the research; F.P. developed the tablet environment and robot setup; R.B. conducted and supervised the data collection, analyzed the data, and wrote the paper; J.V., O.O., S.V., and P.L. supervised; J.V., O.O., S.V., F.P., and P.L. critically reviewed the paper.

Abstract

Previous research shows that the presence of a human peer during a learning task can positively affect children's learning outcomes and enjoyment. However, it is not clear if a robot peer also affects children's learning. The current study aims to find out how second language (L2) vocabulary gains and enjoyment may differ when children follow a vocabulary training: (i) without a peer, (ii) together with a child of the same age, or (iii) together with a humanoid robot. Children were taught six English words in one of these three conditions, to which they were randomly assigned. During the training children were asked to manipulate 3D images of objects on a tablet. Children's word knowledge and enjoyment were measured directly after the training and one week later, via three vocabulary tasks measuring receptive vocabulary and translation of the target words, and questions on enjoyment. Contrary to our expectations, we did not find effects of condition on word learning: Children did not learn more words when they were taught L2 words with a child or robot peer than without a peer, and children without a peer outperformed children in the peer conditions during the delayed post-test. There was no effect of condition on enjoyment, and enjoyment did not impact children's learning outcomes. Thus, a robot peer did not seem to have a benefit for L2 word learning compared to learning without a peer or with a child peer. Future studies could employ more interactive learning tasks in which robots (and human peers) take a more active role in supporting children's learning, and conduct qualitative analyses comparing interactional patterns between child-child and child-robot dyads.

Keywords: child-robot interaction, peer learning, L2 word learning, enjoyment

Humanoid robots have been increasingly incorporated into all kinds of education, including language education. In robot-assisted language learning (RALL) experiments, robots have been employed as tutors (Kennedy et al., 2016; Kory Westlund et al., 2015; Movellan, Eckhardt, Virnes, & Rodriguez, 2009), teaching assistants (Alemi, Meghdari, & Ghazisaedy, 2014; You, Shen, Chang, Liu, & Chen, 2006), and peers (Kanda, Hirano, Eaton, & Ishiguro, 2004; Meiirbekov, Balkibekov, Jalankuzov, & Sandygulova, 2016). The current study is concerned with the latter, thus the use of a robot as a peer in language learning. Previous research shows that the presence of a human peer can positively affect both learning outcomes and enjoyment (King, Staffieri, & Adelgais, 1998; Topping et al., 1997; Yarrow & Topping, 2001). Studies have been using a robot peer to help children learn various skills and topics, such as writing and general world knowledge (Chandra et al., 2018; Hood, Lemaignan, & Dillenbourg, 2015; Okita, Ng-Thow-Hing, & Sarvadevabhatla, 2009). However, no direct comparisons have been made between children engaging in learning tasks with a robot peer versus without a peer or with a child peer. Such comparisons are needed to evaluate the potential for robots in educational settings.

There are several reasons why a peer may enhance learning outcomes. If the peer has more knowledge than the learner, it can engage in the so called learner's zone of proximal development, and help the learner attain the level of the peer. Specifically, peers can transfer their knowledge onto the learner, and thus scaffold the learner's development (Mercer & Littleton, 2007; Rasku-Puttonen, Lerkkanen, Poikkeus, & Siekkinen, 2012; Vygotsky, 1978). If the peer is at a level just above the learner (i.e., in the learner's zone of proximal development; Vygotsky, 1978), they can help the learner attain that level. Another way in which learners may benefit from a peer is by teaching the peer (Rohrbeck, Ginsburg-Block, Fantuzzo, & Miller, 2003). Learning by teaching peers provides learners with the opportunity to practice their knowledge and develop a deeper understanding of the subject. Last, a peer can positively influence task enjoyment. Given the positive relationship between enjoyment and learning outcomes (Gomez, Wu, & Passerini, 2010; Pekrun, Goetz, Titz, & Perry, 2002), this enjoyment may further enhance learning outcomes. Our study, focusing on second language (L2) learning, was aimed at finding out if L2 vocabulary learning and enjoyment differ when children engage in a vocabulary training without a peer, with a child peer, or with a robot peer.

Robots as Peers in RALL Research

Several RALL experiments have used a robot as a peer to help children learn an L2. In Kanda's (2004) seminal RALL study, a robot was placed in classrooms of 119 Japanese-speaking six-year-olds and 109 eleven-year-olds, and children could play with the robot whenever they wanted to. The robot used English words and phrases during these play sessions. Children learned some L2 English words, but only when they continued to play with the robot over the full period of two weeks. Many children, however, stopped playing with the robot within this period and did not learn the words. In another RALL study with multiple sessions, in which three- to five-year-old English-speaking children were taught eight L2 Spanish words over the course of seven sessions in which they played with a robot, limited learning was found as well (Gordon et al., 2016). In fact, children did not exceed chance level on a post-test measuring their learning gains in this study. As can also be concluded from the review presented in Chapter 2, learning gains are generally small.

Other RALL studies with a robot taking the role of a peer consisted of only one session, and found more positive results. For example, preschool children were found to learn L2 verbs better when they made a robot act out these verbs than if they did not teach the robot these words (Tanaka & Matsuzoe, 2012). Thus, learning-by-teaching may be a useful paradigm for RALL. Also, children in an exploratory study were found to learn as many words when playing with a robot as when playing with a child of the same age (Mazzoni & Benvenuti, 2015). A last study shows that children's gender may affect the robot's effects. In an experiment in which children played games with a robot that either always won or lost during those games, children's gender was related to which robot version led to the highest outcomes: Girls learned twice as many words as boys from the ever-winning robot, whereas boys learned twice as many words as girls from the ever-losing robot.

To summarize, some studies show a positive effect of a robot peer on L2 learning, but such an effect has not been found consistently across experiments and may differ between children depending on factors such as gender and the exact role fulfilled by the robot. This may be due to the different types of peer roles that the robot was given in these experiments (e.g., a peer tutor or a peer learner). The first study (Kanda et al., 2004) used the robot as a peer tutor, while the robot was used as a peer learner in the three other studies (Mazzoni & Benvenuti, 2015; Meiirbekov et al., 2016; Tanaka & Matsuzoe, 2012), in which higher learning gains were found. Another explanation may lie in how often children played with the robot. Children could play with the robot over two weeks in Kanda et al. (2004), while they played only once with the robot in Mazzoni & Benvenuti (2015), Meiirbekov et al. (2016), and Tanaka & Matsuzoe (2012).

Learners in RALL experiments enjoy working with a robot (Alemi, Meghdari, & Ghazisaedy, 2015; Han, Jo, Jones, & Jo, 2008; Hsiao, Chang, Lin, & Hsu, 2012; Lee et al., 2011). This may indirectly enhance learning, as studies show that enjoyment enhances learning (Gomez et al., 2010; Pekrun et al., 2002). However, a direct comparison between the effects of children being taught without a peer, together with a child peer, or together with a robot peer on enjoyment in a learning interaction has not yet been made. This has only been investigated in a play setting with older children. In a study by Shahid, Krahmer, and Swerts (2014), eight- and twelve-year-old children played a card-guessing game without a peer, together with a child peer, or together with a humanoid robot. Enjoyment was assessed through self-reports and video observations. Children enjoyed playing with a robot less than playing with a child peer, though more than playing without a peer. It is not clear whether this effect also holds for learning tasks in general and L2 learning in particular. The present study, therefore, will also investigate whether a (robot) peer affects enjoyment and whether this, in turn, affects learning.

The Present Study

The current study aimed to find out to which extent four-to-six year old children learn from and enjoy an L2 vocabulary training when they take this training without a peer, with a child peer, or with a robot peer. Comparing children being taught together with a robot peer to children without a peer enables us to investigate whether a robot peer enhances learning outcomes and enjoyment as compared to being taught without a peer. The additional comparison to a child peer will allow us to investigate whether the presence of a robot peer benefits children’s L2 word learning and enjoyment to the same extent as a child peer.

In our experiment, Dutch children between the ages of four and six years participated in an L2 (English) vocabulary training session. Children were randomly assigned to one of three learning conditions: (i) the no-peer condition, in which they carried out the tasks presented in the training without a peer, (ii) the child-peer condition, in which they carried out these tasks with a child of the same age, or (iii) the robot-peer condition, in which they carried out these tasks together with a humanoid robot. Children’s vocabulary learning gains and enjoyment were assessed immediately after the training. One week later, learning gains were reassessed to measure retention of the target words. The goal of this delayed post-test was to find out whether knowledge decreased, was retained, or even increased over this period of time, due to consolidation processes (see Axelsson, Williams, & Horst, 2016, for a review), and whether learning conditions differentially affected retention of the target words. An increase in word knowledge over a short period of time after the training and immediate post-test is a common finding in word learning experiments (Axelsson et al., 2016). Sleep is one of the reasons for this finding, because it helps strengthen newly acquired information and thus to consolidate new words (Axelsson et al., 2016; Diekelmann, Wilhelm, & Born, 2009; Stickgold & Walker, 2013).

We addressed four research questions:

1. Do Dutch-speaking preschoolers learn more L2 English words in a language game on a tablet when performing this game without a peer, together with a child peer of the same age, or together with a robot?
2. Does performing a language game without a peer, together with a child peer, or together with a robot affect how many L2 words children retain over the period of one week after the training?
3. Do children enjoy performing the language game more when playing without a peer, together with a child peer of the same age, or together with a robot?
4. Does enjoyment mediate the effects of learning condition on children's immediate learning gains and retention?

Based on previous research (King et al., 1998; Topping et al., 1997; Yarrow & Topping, 2001), we hypothesized that children would learn more in both peer conditions compared to when they performed the learning task without a peer. It was not clear whether the highest learning gains should be expected for the robot- or child-peer condition. On the one hand, children in the child-peer condition could be expected to learn most, since children may feel more at ease with another child as compared to a robot, and the robot may distract children (because of its novelty). On the other hand, children in the robot-peer condition could be expected to learn most, as the robot peer took the role of a more-knowledgeable peer (for more information, see the Method section). The robot's utterances were standardized and tailored to the task, and designed to scaffold the child's learning effectively. Peer children's utterances, on the other hand, are more unpredictable and may not scaffold learning to the same extent. Therefore, we tentatively predicted that children in the robot-peer condition would attain the highest learning outcomes, followed by the children in the child-peer condition, and, finally, by the children who completed the training on their own. These effects of condition were expected for both immediate word learning and retention. Given that sleep benefits word learning (Axelsson et al., 2016), leading to higher learning outcomes after a short period of time, we predicted that children would obtain higher scores during the second session that was administered one week later, regardless of condition.

Furthermore, we expected that enjoyment would predict children's word learning gains and retention, independently of condition. Based on Shahid et al.(2014), we expected that children would enjoy the learning task most when they perform the learning task together with another child. This was an additional reason to tentatively predict the highest learning gains in the child-peer condition, given the positive relation between enjoyment and learning outcomes (Gomez et al., 2010; Pekrun et al., 2002).

In the analyses, we controlled for differences in phonological memory, as this is known to be related to vocabulary learning, including L2 vocabulary learning in young children (Gathercole, 2006; Masoura & Gathercole, 2005; Verhagen, Boom, Mulder, De Bree, & Leseman, in press).

Method

Participants

Sixty-seven Dutch preschoolers (26 girls and 41 boys) with an average age of 67.6 months (*SD* = 7.0, range 52 – 78 months) participated. Children were randomly assigned to one of three conditions: without-peer condition (*n* = 23), child-peer condition (*n* = 21) and robot-peer condition (*n* = 23). More information on the characteristics of the children in these three groups is listed in Table 1 in the Data Screening and Analysis section. In addition, 21 children, not included in the analyses, were assigned the role of peer in the child-peer condition. Their word learning was not assessed. Active informed consent from parents/caretakers was obtained for all children. The children were tested at various schools in the Netherlands and had no prior knowledge of English, as indicated in a parent questionnaire. Thirteen children had a different home language in addition to or instead of Dutch, for example, Turkish, Spanish, or Chinese. One additional girl was tested in the child-peer condition, but spoke English at home, and was therefore excluded from the analyses.

Vocabulary training

An L2 vocabulary training was provided by a trained assessor. This training was aimed at teaching six L2 English target words: “heavy”, “light”, “full”, “empty”, “in front of”, and “behind”. These target words were embedded in a narrative, read by the assessor (see Appendix A). Specifically, children were told a story about a zoo that they (and their peer) had to “work in”, by taking care of the animals. Various scenes of animals were displayed on a tablet screen while the story was being read to the children (see Figure 1 for a screenshot of one the scenes displayed on the tablet). Each target word was presented ten times across the narrative. Target children (and their peers) were actively involved in the narrative by having them repeat each target word upon first exposure and answer translation questions ("what is X in English?") on various occasions throughout the narrative. Manipulations of 3D images on a tablet were required at various moments during the training, for example, putting animals in a cage and filling their food trays. In the without-peer condition, children performed all manipulations of the images on the tablet screen. In the peer conditions, the target child and the robot or child peer took turns in performing actions on the tablet. As for repeating and translating target words during the training, the target child, the robot peer, and human peer repeated and translated the words.

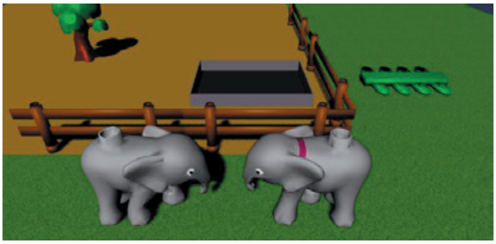


Figure 1. Tablet environment in the vocabulary training. In this scene, children had to put the elephants in their cage and feed them.

Measures

<u>Pre-test</u>

To assess children's knowledge of the six target words prior to the vocabulary training, a pre-test was administered in which children were requested to translate the target words from English to Dutch (e.g., "What is 'light' in Dutch?'). The carrier sentence was in Dutch, and only the target word was said in English. None of the children knew any of the target words, apart from one child who knew the word 'light'. [7]

[7] Rerunning the analyses without this child did not yield different results, so the data from this child was included in the final analyses.

Translation task

To assess children's knowledge of the L2 English word forms, a translation task was used in which children were asked to translate the six target words that had been presented during the training from English to Dutch (e.g., "*Wat is* "light" *in het Nederlands?*" ["What is "light" in Dutch?"]) and from Dutch to English (e.g., "*Wat is "licht" in het Engels?*" ["What is "light" in English?"]). The score was the number of correct answers. Cronbach's alpha showed that the internal consistency of the task was sufficient to good, with α = .71 for the immediate post-test and α = .67 for the delayed post-test. Similar to the pre-test in the translation task from English to Dutch, the carrying sentence was provided in Dutch and only the target word was in English.

Comprehension task

A picture-selection task was used to measure children's receptive knowledge of the target words. In this task, children were presented with four pictures and asked to choose the picture that best represented the target word. For each item, the assessor asked the Dutch equivalent of the question "Where do you see [X]?", with X being the target word in English. For the two prepositions (i.e., "behind", "in front of"), the question was the Dutch equivalent of "Where do you see [X] behind/in front of [Y]?"]. The test consisted of 24 items: four items for each target word. The score was the number of correct answers, yielding a maximum score of 24 points. The internal consistency of the scale was not optimal, with α = .58 for the immediate post-test and α = .66 for the delayed post-test. However, it showed strong correlations with the two translation tasks, so we used a latent factor for the three tasks (for more information, see the section on Data Screening and Analysis).

Enjoyment questions

To measure whether children liked the training, two pictures were shown to them: one of a happy looking boy or girl and one of a sad looking boy or girl (depending on the gender of the child). The assessor then told the children that these children also did the training (either without a peer, with a child peer, or with the robot, depending on the condition that the child was assigned to), and that the happy looking child enjoyed the training, and the sad looking child did not enjoy the training. Subsequently, the experimenter first asked children which boy or girl they resembled most, and then, in a follow-up question, how much they (dis)liked it. Specifically, if children had answered to feel like the happy boy or girl and thus indicated to like the training, the experimenter asked them whether they had liked the training a little or very much. If children had indicated to feel like the sad boy or girl and thus indicated not to have liked the training, she asked them whether they did not like it very much or just a bit. Based on this two-step assessment, enjoyment was rated on a four-point scale ranging from 'did not like it at all' to 'liked it a lot'. We used this assessment rather than a direct question, as young children typically show a yes-bias, perhaps because of their tendency to give socially desirable answers (Moriguchi, Okanda, & Itakura, 2008).

Attitude towards the robot and perception of its role

Children in the robot-peer condition were asked how much they liked the robot in a similar fashion to the general enjoyment question, resulting in a four-point scale ranging from 'The child did not like the robot at all' to 'The child liked it a lot'. We asked additional questions about how children perceived the robot via four forced-choice questions. Children were asked 1) whether they perceived the robot as a friend or a teacher; 2) as a teacher or an object; 3) as an object or a friend; 4) as a human or an object. These questions were used to control whether our framing of the robot as a peer had succeeded.

Nonword repetition task

Phonological memory was assessed with a nonword repetition task. This task contained a sub-set of the items used in Rispens and Baker (2012), embedded in a computerized task appropriate for young children (for more details, see Verhagen, de Bree, Mulder, & Leseman, 2017). In this task, children were instructed to repeat pre-recorded, non-existing words. The test phase contained twelve items, preceded by two practice items. The answers were scored online by the assessor and rewarded one point for each correctly repeated word, so the maximum score on this task was twelve. Cronbach's alpha showed that the reliability of this task was just below acceptable, α = .63, but α = .83 in the original paper (Verhagen, de Bree, et al., 2017). Ten percent of the data was scored by an additional researcher, showing good interrater agreement with 86% agreement, κ = .71 (95% CI, .562 to .864), p < .001.

Robot

The robot used in the present study was a Softbank NAO robot, a 58cm tall robot with a humanoid appearance. The study used the Wizard-of-Oz approach, with an experimenter – who was a different person than the assessor providing the vocabulary training – controlling the robot and the flow of the interaction during the training. This controlling was done via a graphical user interface from a laptop computer located in the experiment room but not in direct sight of the child. The robot's responses had been preprogrammed, such that its responses and behaviors were consistent for all children. However, there was also the possibility to type in text online and adapt the robot's utterances to the child's utterances. For example, one child indicated that they did not believe that the robot could see them, so the experimenter had the robot say, "I can see you" to stimulate the interaction between the robot and the child and give it a personal feel. However, such utterances were kept to a minimum to ensure that interactions were comparable across children.

During the vocabulary training, the robot was sitting in crouch position and performed several behaviors. Those behaviors were: 1) manipulating (e.g., swiping, tapping) the images on the tablet by moving its arm; 2) repeating the target words; 3) commenting on the children's manipulations (e.g., "yes, now the elephant is in its cage"); 4) pointing to the tablet while explaining what to do, in case children failed a task. Types 1 and 2 behaviors were the same activities as the target child (and child peer) was/were asked to do, type 3 behavior was included to increase children's motivation and stimulate interaction, and type 4 behavior was used for scaffolding. If children performed the manipulation correctly (e.g., they put all food in the food tray), the robot made a general comment (e.g., "wow, so much food"). If children performed the manipulation incorrectly or did not know what to do (e.g., did not know where to put the food), the robot made a helpful comment and a pointing gesture (e.g., "I think it has to go over there" while pointing at the food tray). Likewise, when answering the translation questions ("what is X in English?"), the robot would either say "yes, that is X" or "I think it is X", depending on whether children knew the answer. In other words, the robot acted as a slightly more knowledgeable peer during the vocabulary training.

Procedure

Prior to the individual test sessions, all children participated in a group demonstration of the robot in which the robot introduced itself and did a dance with the children. Subsequently, the children were tested individually in a quiet room in their schools. The first session started with the pre-test (the translation task from English to Dutch) to assess whether the children had any prior knowledge of the target words. If children had been assigned to the child-peer condition, the child peer entered the room after the target child had completed the pre-test. If children had been assigned to the robot-peer condition, the robot was already present in the room during the pre-test, but only "woke up" once the target child had completed the pre-test. See Figure 2 for the setup.

Prior to the vocabulary training, children (and their peers) were invited to play with the tablet, to allow them to practice with the tablet and to make sure they would focus on the contents of the training rather than the tablet itself. Then the vocabulary training was provided. Once the training had been completed, the peer left the room and the target child completed the test battery individually. The order of the test battery was as follows: enjoyment questions, picture-selection task, the translation task from English to Dutch, and the translation task from Dutch to English. Two additional tasks were administered but will not be discussed in this chapter, as the test-retest reliability of one measure was inadequate, (*r*(64) = .09, *p* = .497) and the other measure did not assess children's word learning, and is, therefore, beyond the scope of the current study. One week after the training session, children's retention was measured at a second post-test. In this delayed post-test, all measures, except the enjoyment questions, were administered again in the same order, and the nonword repetition task was administered as the final task. For each task completed, children received a sticker. At the end of each session, they got a small gift as a reward for their participation. The first session lasted about 50 minutes, and the second session about 30 minutes.

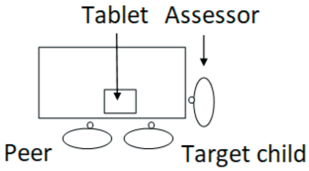


Figure 2. Experimental setup in the peer conditions.

Data Screening and Analysis

Data preparation and missing data

Pearson's correlation tests showed that the translation tasks and comprehension task were highly correlated during both post-tests, *r* (64-66) = .384-.657, all *p*s < .002.

Therefore, children's scores on the three tasks measuring word learning (i.e., the two translation tasks and the comprehension task) were combined using a longitudinal Confirmatory Factor Analysis[8]. Using this method, the scores of the tasks were combined into a single latent structure while maintaining the factorial invariance, using the statistical program Mplus (Muthén & Muthén, 2010). That is, the factor loadings of the three tasks were constrained to be equal between the two post-tests and the mean of the second post-test was left free, while the mean the of the first post-test was set to zero (see Little, 2013). The resulting scores on the two latent factors (one for each post-test) could then be compared, as they were scaled in an identical manner. Given the sample size, the research questions could not be tested using a model containing these latent variables. Therefore, exported factor scores were used in further analyses. Word learning scores on all tasks were missing for one child in the immediate post-test and for two children in the delayed post-test due to illness. The data on the enjoyment measure indicated that even though we administered a four-point scale, only two of the scale options were used: All children indicated to like the training either a little or a lot. Therefore, the enjoyment variable was dichotomous. Enjoyment data of nine children were missing due to an error in the test protocol.

Group characteristics

We performed several analyses to inspect the data before we analyzed it to address our research questions. First, we checked whether the three groups (i.e., without a peer, together with a child peer, or together with a robot peer) did not differ in gender, age, and phonological memory. A Pearson's Chi-Square test was carried out with gender

[8] Note that this approach differs from the one in Chapter 3, in which we did not combine the task scores despite similar correlations. The reason for this difference is that, in Chapter 3, we were interested in possible differential effects of condition on type of word knowledge (receptive knowledge and translation abilities) gained during the vocabulary training.

and condition as categorical variables to investigate the gender composition of the three groups. There was no significant difference in gender between the three groups, χ^2 (2, *N* = 67) = 2.39, *p* = .304, φ = .189. A one-way ANOVA showed a significant difference in age across the three groups, (*F*(2, 64) = 3.21, *p* = .047, η_p^2 = .09). A post-hoc Tukey test showed that the children who took the training without a peer were significantly older than children in the robot-peer condition (*p* = .050), but not significantly older than children in the child-peer condition (*p* = .155). Similarly, a one-way ANOVA indicated that there was a significant difference across the three groups in phonological memory, (*F*(2, 62) = 3.419, *p* = .039, η_p^2 = .10). Post-hoc Tukey tests showed that children in the robot-peer condition had significantly higher phonological memory scores than children in the without-peer condition (*p* = .038), but not than those in the child-peer condition (*p* = .221). Thus, we included age and phonological memory as covariates in the subsequent analyses.

Table 1. Gender Composition, Age, and Phonological Memory Scores of the Without-Peer, Child-Peer, and Robot-Peer Groups

	Without peer	Child peer	Robot peer
	(*n* = 23)	(*n* = 21)	(*n* = 23)
n girls (%)	10 (43%)	5 (24%)	10 (43%)
M age in months *(SD)*	70.48 (4.25)	66.67 (7.68)	65.70 (7.78)
M phonological memory *(SD)*	6.65 (1.77)	7.84 (2.29)	8.22 (2.26)

Analyses

To investigate whether children learned from our training, we compared children's scores on the comprehension task against chance level (25%). Since there was no chance-level performance in the translation tasks, we compared children's scores on the translation tasks to a score of zero. We compared children's task score against zero or chance level separately for each condition, and thus applied a Bonferroni correction. To address our first and second research question on the effects of learning condition on L2 word learning and retention, we ran a repeated-measures ANCOVA. The dependent variable was the word-learning score (the factor score of the three word-learning tasks) during the immediate and delayed post-test, the independent variable was condition, and covariates were age and phonological memory. To address our third research question on the effects of learning condition on children’s enjoyment of the training, a Pearson’s Chi-Square Test was carried out with enjoyment (‘enjoyed a little’ vs. ‘enjoyed a lot’) and condition (without peer, child peer, or robot peer) as categorical variables. To address our fourth and last research question on whether enjoyment mediated effects of learning condition on children’s word learning and retention, we re-ran the repeated-measures ANCOVA with ‘enjoyment’ as an additional independent variable.

Results

Results on Word Learning and Retention

Descriptive statistics for the word learning tasks during the first and second session can be found in Table 2. Note that these are true task scores rather than the exported factor scores that we used in the subsequent analysis. Irrespective of condition, children performed significantly above chance level and zero performance during the immediate and the delayed post-test (see Table 3 for the exact results of these analyses).

As outlined above, children’s scores on the three tasks were combined for each post-test using a longitudinal confirmatory factor analysis. The resulting model fitted the data well ($\chi^2(12) = 12.99$, $p = .370$, RMSEA = .04, CFI =.99, TLI = .99; see Appendix B for the full model). Factor loadings were all significant and high (range .63-.81).

Table 2. Means (Standard Deviations) of the True Task Scores and the Exported Factor Scores for the Three Tasks during Both Post-Tests

		Without peer	Robot	Child
English to Dutch	Immediate	1.23 (1.34)	1.30 (1.11)	1.48 (1.12)
	Delayed	2.17 (1.56)	1.83 (1.44)	1.89 (0.99)
Dutch to English	Immediate	1.05 (1.29)	1.26 (1.29)	0.67 (1.07)
	Delayed	1.74 (1.25)	1.35 (1.07)	1.26 (0.99)
Comprehension	Immediate	12.27 (3.78)	11.78 (2.68)	11.67 (3.37)
	Delayed	14.30 (3.44)	12.61 (3.81)	12.11 (2.75)
Factor scores	Immediate	0.34 (2.43)	-0.12 (2.12)	-0.24 (1.72)
	Delayed	1.86 (2.86)	1.20 (2.53)	1.08 (2.00)

Note. The maximum score was six for the translation tasks and 24 for the comprehension task. The scale of the factor scores is arbitrary.

Table 3. One-Sample T-Tests against Zero (for the Translation Tasks) and against Chance-Level Performance (for the Comprehension Task)

		Without peer		Robot peer		Child peer	
		t(df)	*d*	*t*(df)	*d*	*t*(df)	*d*
English to Dutch	Immediate	4.29(21)***	.30	5.66(22)***	.66	6.02(20)***	1.86
	Delayed	6.70(22)***	.97	6.10(22)***	.80	8.31(18)***	.69
Dutch to English	Immediate	3.80(21)**	.15	4.70(22)***	.38	2.87(20)**	.89
	Delayed	6.67(22)***	.97	6.04(22)***	.78	5.56(18)***	.80
Comprehension	Immediate	7.78(21)***	.34	10.35(22)***	3.05	7.71(20)***	.38
	Delayed	11.56(22)***	.41	8.32(22)***	.45	9.69(18)***	.15

Note. ** < p < .01; *** p < .001.

A repeated-measures ANCOVA indicated no significant effects of condition, time, or an interaction effect between condition and time (see Table 4 for the results). Trends were found for the covariates age and phonological memory: Older children and children with better phonological memory skills tended to perform better than younger children and children with poorer phonological memory skills.

Table 4. Results of the Repeated-Measures ANCOVA on Effects of Learning Condition on Word Learning and Retention

	F (df)	*p*	η_p^2
Condition	.69 (2, 60)	.506	.02
Time	.08 (1, 60)	.776	.00
Condition * time	1.91 (2, 60)	.157	.06
Age	3.59 (1, 60)	.063	.06
Phonological memory	3.53 (1, 60)	.065	.06

Results on Enjoyment, Role of Robot, and Mediating Effects of Enjoyment

Chi-Square Tests revealed no significant differences in enjoyment between the three groups, χ^2 (2, N = 61) = 4.58, p = .101, φ = .274. Eighteen out of the 23 children in the robot-peer condition indicated to perceive the robot as a friend rather than as a teacher, indicating that our framing of the robot as a peer succeeded.

Re-running the repeated-measures ANCOVA with enjoyment as an additional independent variable yielded different results than the model without enjoyment (see Table 5 for the exact results of these analyses). Specifically, this analysis showed an interaction effect between condition and time, which indicated that children in the without-peer condition improved slightly more from the immediate to the delayed post-test than children in the peer conditions (see Figure 3). Confidence intervals showed that the performance of children in all three conditions did not significantly differ from the mean during the immediate post-test (95% CI without-peer condition [-0.16, 3.08], child-peer condition [-1.29, 0.64], robot-peer condition [-1.23, 0.83]. During the delayed post-test, only children in the without-peer condition performed significantly higher than the mean (95% CI without-peer condition [1.31, 5.14], child-peer condition [-0.17, 2.11], robot-peer condition [-0.09, 2.35]).

Table 5. Results of the Repeated-Measures ANCOVA on (Mediating Effects of) Enjoyment

	F (df)	*p*	η_p^2
Enjoyment	.79 (1, 51)	.380	.02
Condition	1.96 (2, 51)	.152	.07
Time	.36 (1, 51)	.553	.01
Enjoyment * condition	.88 (2, 51)	.419	.03
Enjoyment * time	1.71 (1, 51)	.197	.03
Condition * time	3.33 (2, 51)	.044	.12
Age	2.32 (1, 51)	.134	.04
Phonological memory	3.04 (1, 51)	.087	.06

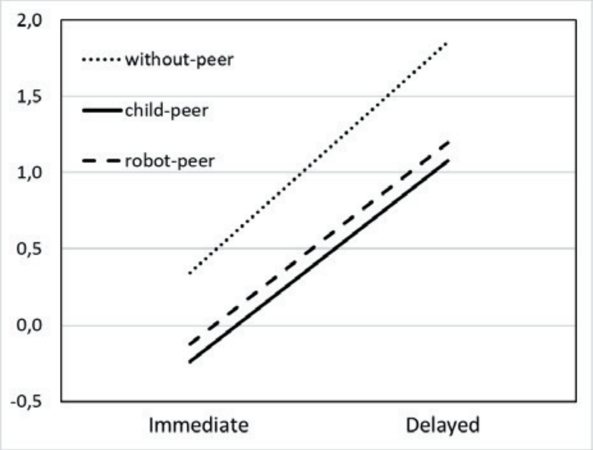


Figure 3. Condition * time interaction.

Discussion

The main research question of this study was whether learning of L2 English words and enjoyment of the vocabulary training by 4- and 5-year-old Dutch preschoolers differed between children taking the training without a peer, together with a child peer of the same age, or together with a robot peer. We expected children to learn more in both peer conditions than in the without-peer condition. There were competing hypotheses regarding the type of peer that would have the greatest benefit: the child peer or the robot peer. A robot peer may provide the most effective scaffolding, but the child peer may lead to higher learning gains through higher enjoyment of the task (cf. Shahid et al., 2014) or through making the child feel more at ease. We also expected children to enjoy learning most with a child peer, followed by a robot peer, and least without a peer. Finally, we predicted that this increased enjoyment, in turn, would lead to higher word learning gains.

Overall, children learned from our training, as they performed above chance level on a word comprehension task and above zero in English- to Dutch and Dutch-to English translation tasks. Across conditions, children obtained higher scores on the comprehension task than the translation tasks, in line with previous research showing that vocabulary is much harder to learn productively than receptively (Mondria & Wiersma, 2004). Children were expected to show higher performance on the delayed post-test than on the immediate post-test, but no significant differences were found between the two post-tests. Contrary to our expectations, taking the training together with a peer, either human or robot, did not benefit word learning over taking the training without a peer. There even was a benefit on retainment of taking the training without a peer, but this effect was only found in the model that included enjoyment and should therefore be interpreted with caution. Last, enjoyment did not significantly differ between conditions and did not moderate the effects of condition on learning, in contrast to our expectations.

A possible explanation for the lack of peer benefits in our study is that the robot and child peers actually distracted the target children, rather than helped them to learn the L2 words. Previous work has shown that robots can distract children during a learning task, in particular when they show too much social behavior (Kennedy et al., 2015b). Another explanation is that the vocabulary training did not allow for enough interaction between the learner and the peer for the learner to benefit from the peer. Since the learner and the peer had to take turns in manipulating the images on the tablet, each had fewer opportunities to manipulate images on the tablet than children taking the training without a peer. Specifically, learners in the peer conditions only got half of the opportunities to manipulate the images on the tablet as compared to the learners without a peer. We tried to stimulate the interaction between the robot and the child and to have the child learn from the robot by having the robot respond contingently to the child's actions, but perhaps more interaction is required (cf. Webb, 1989) between the learner and the robot or child peer for the learner to truly benefit from the peer. Furthermore, the robot was controlled via an assistant in real time, which entailed that its responses were rather slow. This may have led the robot to appear not highly engaged in the interaction. Another potential explanation for the lack of effects is that there was another person involved in the interaction, as there was an assessor present providing the vocabulary training. In our experiment, target children did not have to rely on their peers to complete the tasks, they could also turn to the assessor for help, which, however, occurred only occasionally. Last, for a robot to truly be an effective peer, it should provide input and react to the child's behaviors within the child's zone of proximal development and scaffold the learner's development by transferring knowledge at a level just above that of the learner (Mercer & Littleton, 2007; Rasku-Puttonen et al., 2012; Vygotsky, 1978). Our robot could not adapt to the learner's level of knowledge. Perhaps, robots are more effective peers if they are adaptive, which is a line for future research. Previous research on adaptive robots has shown that adaptivity may benefit learning speed and engagement (Schodde, Bergmann, & Kopp, 2017; de Wit et al., 2018).

Note that our setting may not have been ideal to stimulate peer benefits, as discussed above, but peer benefits should be expected anyway. Children in the peer conditions had more exposure to the target words than children in the without-peer condition, as they also heard the target words from the peers, who were also required to answer the translation questions as well. Thus, when practicing the target words, the robot and the human peer also pronounced these words, and target children received additional input. Given the lack of peer benefits, however, this additional input did not outweigh (negative) effects of the peers (e.g., a distraction or less time to manipulate objects on the tablet).

Furthermore, we found no effect of condition on children's self-reported enjoyment, also contrary to our expectations. Differences in enjoyment ratings were not related to learning success. Our results for enjoyment should be interpreted with caution, as it is difficult to measure enjoyment in young children. Young children typically show a yes-bias, leading them to give socially desirable answers (Moriguchi et al., 2008). While this mostly holds for younger children, enjoyment and liking are still difficult to measure in preschoolers. We tried to diminish this bias by showing two pictures when assessing enjoyment: one picture of a boy/girl that did not like the training, and one picture of a boy/girl that did like the training. We do not know whether this solved the problem, given that none of the children indicated to not have liked the training. So, on the basis of the present data, we cannot tell whether children were biased in their responses or whether their answers were reliable and they truly enjoyed the game to some extent. Another option is that the scale we used was not sensitive enough to pick up on more fine-grained differences in children's enjoyment.

Last, we found that children obtained higher scores at the delayed post-test than at the immediate post-test. This finding is in line with previous research (Axelsson et al., 2016). Children need to integrate the new words into their existing memory. Sleep helps in strengthening this knowledge (Diekelmann et al., 2009; Stickgold & Walker, 2013). Most research on consolidation is on L1 learning, and our findings show that similar processes may come into play when learning L2 words. Furthermore, we found that children obtained higher scores if they had higher scores on the nonword repetition task measuring phonological memory, also in line with previous research (Gathercole, 2006; Masoura & Gathercole, 2005; Verhagen, Boom et al., in press). This result suggests that children who are better able to repeat nonwords are better able to store sounds in their short-term memory, which is a useful skill for learning new words (Cheung, 1996; Gathercole, 2006; Gathercole & Baddeley, 1990; Service, 1992).

A possible concern in RALL research has been whether robots that were found to be effective, were effective due to the way in which they were used, or due to learning gains being boosted by the novelty of the robot. If the latter were the case, robots' effects would ware off once children see the robot more than once. Our results do show that even when a robot is presented for the first time, it does not necessarily boost learning gains more than when children are taught together with a human peer or without a peer. Perhaps, in our study, there were two opposite effects that canceled each other out: The novelty of the robot may have facilitated learning, but the robot's limitations discussed above (i.e., slow responses, lack of adaptivity, fewer opportunities for target children to manipulate objects on the tablet) may have canceled out possible novelty benefits.

Our experiment has several limitations. First, our interaction was restricted in the sense that children could not freely play with their peer and explore whether they could learn from each other. It is a challenge to design such free and explorative child-robot interactions given the current state of technology. Robots should have more advanced abilities to perceive their environment to engage in non-structured and explorative interactions. Furthermore, there was a human teacher present, who affected the interaction. Despite these limitations, this study was the first to make the comparison between children performing a language-learning game without a peer, with a child peer, and with a robot peer, and provides, therefore, useful insights into how being taught together with a robot compares to other learning situations. Also, we thoroughly assessed children's word learning by using multiple word-learning tasks assessing different types of vocabulary knowledge, and administering them at two time points. The delayed post-test allowed us to investigate not only how a (robot) peer may affect direct learning, but also retention over a short period.

Future research could look into more interactive learning tasks, to see whether a robot peer may result in higher learning outcomes when there is more interaction between the robot and the learner. Also, tasks in which no other parties are involved may stimulate the interaction between the learner and the robot. Interactive learning tasks consisting of more than one session could also be informative to the question on how taking a training with a robot peer compares to taking a training without a peer or with a child peer. Perhaps familiarity with the robot could affect the degree to which children benefit from working together with it. In previous RALL studies involving multiple sessions (Gordon et al., 2016; Kanda et al., 2004), limited learning gains were found. However, in these studies, the robot took the role of a peer-tutor rather than that of a peer-learner, and the robot was not directly compared to a child peer or the absence of a peer. So, it as yet an open question whether robot peers benefit learning over multiple learning sessions. More elaborate direct comparisons of robot peers to human peers are necessary and informative for the further development of humanoid robots, but the finding that a robot peer does not differ from a human peer gives hope for future research on robot peers in education. Future research could look further into the analysis of interactional patterns during learning tasks similar to the one used in this experiment. Comparisons of interactional patterns between children being taught together with a child peer and children being taught together with a robot peer (e.g., the degree to which they look at the peer, talk to the peer, or follow up on suggestions provided by the peer) could stimulate further development of humanoid robots, and their implementation in language learning environments. In our study, we cannot disentangle whether the robot's behavior or the robot itself led to a lack of peer benefits. Studying interactional patterns and the robot's behavior may help investigate this issue further.

To summarize, we did not find the expected advantage of a peer on L2 word learning gains or enjoyment over children taking a vocabulary training without a peer. Future studies with more interactive vocabulary trainings and other ways to measure enjoyment with young children should be carried out to further investigate the effect of a child or robot peer on L2 word learning.

Appendix A

This appendix contains a translated version the narrative of the vocabulary training. The original narrative was in Dutch. Only the words in **bold** were said in English.

Narrative

Look, there is a zoo over here. One of the caretakers of the animals has fallen ill, so he cannot make it to the zoo today. You are asked to replace him. That means you will take care of the animals today. That is fun!

The zoo is in England. In England, people speak a different language from Dutch, namely: English. You will also learn some English words today. That is useful, because then you will know the language of the zoo a bit. Listen closely to the story and we will see later which English words you learned!

First, we are going to check on the elephants. Look, here are a mommy and a daddy elephant. The mommy elephant is the elephant with the pink bow. The elephants need to be put in their cage. The mommy elephant has a baby in her belly, so she is very heavy. The English word for “heavy” is “**heavy**” {If the child does not repeat the word, say: “Repeat after me: “**heavy”**}. Very good: **heavy**! Put the “**heavy**” mommy elephant carefully on the other side of the fence, inside the cage. What is heavy in English? {If the child does not say the target word, say: Heavy is in English “**heavy**”}. Well done! The daddy elephant does not have a baby in its belly of course. He is not very “**heavy**”. The daddy elephant is light. In English “light” is **“light”** {If the child does not repeat the word, say: “Repeat after me: “**light”}**. Well done: **light**! Put the “**light”** daddy elephant next to the “**heavy”** elephant in the cage, behind the fence. What is light in English? {If the child does not say the target word, say: Light is in English “**light**”}. Well done, now the “**heavy**” mommy elephant and the “**light**” daddy elephant are in their cage.

Can you check whether there is enough food in their tray? Oh it is completely empty. The English word for empty is “**empty**”. {If the child does not repeat the word, say: “Repeat after me: “**empty”**}. Good job: **empty**! The food tray is “**empty**”. What is empty in English? {If the child does not say the target word, say: Empty is in English “**empty**”}. You can put the sticks with leaves in the tray. Now the tray no longer is “**empty**”. Now it is full. In English that is “**full**”. {If the child does not repeat the word, say: “Repeat after me: “**full”}**. Yes, good job: **“full”**! What is full in English? {If the child does not say the target word, say: Full is in English “**full**”}. Well done!

Now we are going to the polar bears. There is a big mommy polar bear and a little one. Can you take the mommy polar bear out of the water and in front of the boat? The mommy polar bear is **heavy.** The little polar bear cannot stay in the water by itself. Can you put it next to its mommy? The baby polar bear is not “**heavy**”; it is “**light**”.

The polar bears are now in front of the boat. The English word for in front of “**in front of**”. {If the child does not repeat the word, say: “Repeat after me: **“in front of”**}. Good job, **in front of!** So what is “in front of” in English? {If the child does not say the target word, say: In front of is in English “**in front of**”}. There is a bucket with fish “**behind**”. **Behind** means “behind”. {If the child does not repeat the word, say: “Repeat after me: “**behind**”}. Good job, **behind!** The bucket is now full, because there is a very big fish in it: it is **“full”.** You can give the fish to the polar bears. The bucket is now empty, it is “**empty**”. Put the **“empty”** bucket back “**behind**” the boat. What is “behind” in English? {If the child does not say the target word, say: Behind is in English “**behind**”}. You can also clean up the pool, because it was rather “**full**” with the polar bears and all the things. Put the ball into the box in front of the tree, so the box “**in front of”** the tree. Good job, now the box is “**full**”. The watering can is completely empty, so it is very light to lift, very “**light”**. Put the watering can “**in front of”** the tree as well. You did a really good job! The pool is empty now, completely “**empty**”.

Look, now there are giraffes! Look at that, some of the giraffes are playing hide and seek with us! Do you see where they are? There is one giraffe over there, behind those trees! It is “**behind**” the trees. It was well hidden “**behind**” the trees. There is also a giraffe sleeping in front of the bushes. It lies **in front of** the bushes. Can you put both giraffes “**in front of”** their cage, so in front of the cage? Well done! Now you can give them some food. The food tray is still in the corner. It is **behind** the bush, so behind the bush. There is nothing in it yet, it is completely “**empty**”. It is still “**light”**. Put it “**in front of**” the giraffes, then it is easy for them to reach it while eating. Can you fill the food tray with leaves? The leaves are also very “**light**”. Good job! Now the tray no longer is **“empty”,** it is completely **“full”.**

It is evening! You can bring the polar bears to their sleep pool. Put the “**heavy**” mommy polar bear behind the fence, so **“behind”** the fence, into the pool. Good job! Put the baby polar bear also “**behind**” the fence into the pool. The polar bears get to play for a little while before they go to sleep. Put the ball also in the pool. Put it “**in front of**” the mommy polar bear, so in front of the mommy polar bear. The ball is very light, very “**light”**, and otherwise the wind may blow it away! Well done! The pool is now “**full**”, so full, with two polar bears and a ball. The food tray is “**in front of**” the pool. Can you put the sticks with leafs also in the food tray? Then it is “**full**”, in case the polar bears want to eat something tonight. Good job! Now the cage is completely “**full”** and they can go to sleep.

Wow you did such a good job! I will tell the caretaker tomorrow everything that you did. You fed the “**heavy**” mommy elephant and the “**light**” daddy elephant. Their food tray was completely “**empty**” and you made sure it was “**full**” again. You also put the polar bears “**in front of**” the boat so that you could feed them a fish. Then the giraffes were playing hide and seek, one of them was “**behind**” the trees, and you fed them. And you also brought the polar bears to bed. Very good job!

Appendix B

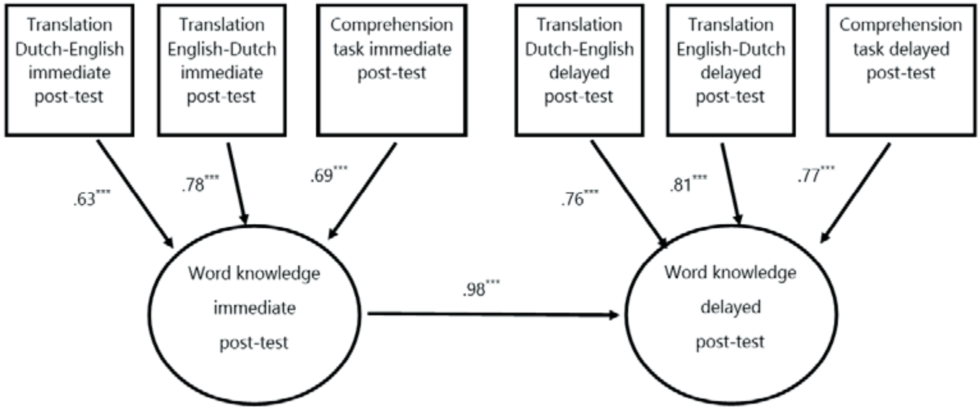

Translation Dutch-English immediate post-test
Translation English-Dutch immediate post-test
Comprehension task immediate post-test
Translation Dutch-English delayed post-test
Translation English-Dutch delayed post-test
Comprehension task delayed post-test
.63***
.78***
.69***
.76***
.81***
.77***
Word knowledge immediate post-test
.98***
Word knowledge delayed post-test

Figure B.1. Model of the confirmatory factor analysis.

5

Individual differences in children’s language and attention skills as moderators of robot-assisted language learning

Rianne van den Berghe, Ora Oudgenoeg-Paz, Josje Verhagen, Susanne Brouwer, Paul Leseman, & the L2TOR team*. (in preparation).

* The L2TOR team consists of Mirjam de Haas, Jan de Wit, Bram Willemsen, Emiel Krahmer, Paul Vogt (who conceptualized the research and collected the data), Thorsten Schodde, Cansu Oranç, Junko Kanero, and Chris Wallbridge (who designed the lessons and the L2TOR setup). The L2TOR project is led by Tony Belpaeme (University of Ghent) and Paul Vogt (Tilburg University). PIs are Tony Belpaeme, Paul Vogt, Emiel Krahmer, Stefan Kopp, Aylin Küntay, and Paul Leseman.

Author contributions: R.B., O.O., J.V., and P.L. conceptualized the research; R.B. and O.O. collected the data, R.B. wrote the paper; S.B. analyzed the data; O.O., J.V., and P.L. supervised; O.O., J.V., S.B., and P.L. critically reviewed the paper.

Abstract

The current study investigated the added value of social robots for teaching L2 vocabulary to young children, while specifically focusing on the moderating role of individual child characteristics deemed relevant for language learning. So called tier two words in the domains of mathematical and spatial language were taught in an L2 English vocabulary training intervention consisting of seven sessions over a three- to four-weeks period. This training was administered to 193 Dutch children, who were randomly assigned within schools to one of three experimental conditions: (a) with a tablet only, (b) with a tablet and a robot that used deictic (pointing) gestures, or (c) with a tablet and a robot that used both deictic and iconic gestures (gestures depicting the target word), or to a random control condition in which children did not receive a vocabulary training but played dancing games with the robot as a placebo instead. To assess word learning, we used translation tasks and a receptive word comprehension task. As potential moderator variables, we measured children's L1 vocabulary knowledge, phonological short term memory, and selective attention. Children in the experimental conditions were found to outperform children in the control condition on all L2 word-knowledge tasks at an immediate and delayed post-test, respectively. However, there were no differences between the experimental conditions. Thus, children did not benefit more from the robot or from the robot with iconic gestures next to a tablet compared to the tablet-only condition. Moderating effects were found for the language and attention skills. Together, the results show that robots, given the current state of the technology, are not more effective than the cheaper and more accessible tablet technology. The results, furthermore, highlight that taking individual differences in language and attention skills into account is highly relevant in designing and evaluating robot-assisted L2 learning.

Keywords: educational robots, second-language learning, child-robot interaction, individual differences, language and attention skills

Overview

The current study addresses the use of social robots in language education (see Chapter 2; Kanero et al., 2018, for reviews). Several potential advantages of robot-assisted language learning (RALL) have been identified in the extant research relative to language learning assisted by traditional technologies such as tablets. Social robots allow for interactions that make use of the physical environment (e.g., acting upon objects, enacting particular movements or operations, using various types of gestures) and they can stimulate more natural, human-like interactions because of their humanoid appearance. Current evidence on the effectiveness of RALL, however, is mixed (Chapter 2; Kanero et al., 2018), indicating that there may be no overall advantage of social robots as tutors in language learning. Moreover, a study has found that RALL was only effective for a subgroup of children (Kanda, Hirano, Eaton, & Ishiguro, 2004), suggesting that individual characteristics of children may moderate the effects of RALL. It is possible that robots are useful language-education tools for certain children only, for example, depending on children's prior language knowledge and general (language) learning abilities. However, studies into the role of individual child characteristics in RALL, enabling the identification of such specific groups, are scarce. The current study, therefore, aims to add to the evidence regarding the effectiveness of robots in second language (L2) teaching of young children, while specifically focusing on the role of individual differences in relevant skills, in particular children's first language (L1) vocabulary knowledge, their phonological short term memory capacity, and their selective attention skills.

Robot-Assisted Vocabulary Learning

For RALL to be of added value to (second) language learning in education settings, that is, to have wider impact relative to other technology, but also to human tutoring, RALL should be capable of teaching children not just a few, but a substantial number of words. This requires a series of sessions over a longer period, as is common practice in effective and impactful vocabulary training programs involving human tutors (Marulis & Neuman, 2010). A critical and counter-intuitive finding in this regard, however, is that robot-assisted teaching of a few L2 words in a single session (de Wit et al., 2018; Tanaka & Matsuzoe, 2012) has been found to be more effective than teaching children words over multiple robot-assisted sessions (Gordon et al., 2016; Kanda et al., 2004).

A possible explanation for this counter-intuitive finding may be the novelty of the robot. Since children generally have little or no experience with robots, they may attend more to the robot and become more motivated by it, and thus learn more, than when they would have been more familiar with robots (see Leite, Martinho, & Paiva, 2013, for an overview of long-term interactions with robots). Hence, it is possible that when children work together and interact with the robot over multiple sessions, which would be required to teach them a larger vocabulary, the increasing familiarity with the robot will lead to decreasing learning in later sessions. To rule out a short-lived novelty effect as a main cause of increased word learning in single sessions, multiple-session studies are required. Novelty effects are often confounded with the true effects of technology-enhanced L2 education or lack thereof (Salaberry, 2001), and there is a distinct risk of overestimating the potential of technology-enhanced education when applied to real education settings. The present study, therefore, evaluates RALL of a relatively large number of words in a multiple-sessions design.

In addition, while studying the effects of robots is, in itself, important in view of applications in education, it is crucial to compare the effectivity of robots to that of other, possibly cheaper and more accessible technological aids such as tablets. Previous studies did not provide a clear answer to the question whether children learn more in robot-assisted lessons as compared to lessons in which other types of technology are used. Only one study has directly compared a human teacher, a robot and a tablet in a one-session vocabulary game and found no difference in learning gains (Kory Westlund et al., 2015). Some RALL experiments, where children were taught reading skills, suggest that children learn more from a robot than a tablet (Gordon, Breazeal, & Engel, 2015; Han, Jo, Jones, & Jo, 2008; Hyun, Kim, Jang, & Park, 2008), but the overall evidence is mixed at best, including earlier findings on more general skills, such as puzzle-solving (Leyzberg, Spaulding, Toneva, & Scassellati, 2012). Thus, the few, mainly single-session studies do not provide conclusive evidence yet on the possible benefits of robots over other forms of technology when applied in a more realistic multiple-sessions word learning program.

Individual Differences in L2 Learning

Learning an L2 is dependent on both the quality and quantity of the L2 input (Hoff, 2013; Unsworth, 2016) and on characteristics of the learner (i.e., the learner’s cognitive and personality resources; Cummins, 1991). Specifically, the learner’s prior L1 vocabulary knowledge may benefit L2 word learning, including robot-supported L2 learning. Thus, in addition to a main effect of RALL, children with larger L1 vocabularies may benefit even more from RALL than children with smaller vocabularies, because they can use their richer lexical and conceptual networks to disambiguate new input and to integrate it in existing knowledge. This phenomenon, found in particular for reading instruction but also in vocabulary learning (e.g., Penno, Wilkinson, & Moore, 2002), is referred to as the Matthew effect (Stanovich, 2009). The role of prior L1 knowledge, however, may be less straightforward in L2 learning (Wolter, 2006). For concepts that are similar in the L1 and L2, the learner can simply map the new L2 label onto the underlying concept. However, when conceptual systems are different between L1 and L2, simple mapping may not be possible or easily lead to a wrong mapping of L2 labels onto L1 concepts, requiring more thorough restructuring of the conceptual knowledge (Hemsley, Holm, & Dodd, 2013; Wolter, 2006). Besides the conceptual similarity, similarity in word form between L1 and L2 can also aid in L2 learning, at least when this similarity also entails similarity in meaning (Brenders, van Hell, & Dijkstra, 2011; Hemsley et al., 2013; Sheng, Lam, Cruz, & Folton, 2016). In the current study Dutch speaking children were taught English words. As English and Dutch are both Germanic languages with highly similar conceptual systems and many similarities in word form, we expected that a larger L1 vocabulary would help children in learning English words. Thus, individual differences in L1 vocabulary were expected to moderate the effect of robot-supported L2 instruction.

Another factor that may moderate the effect of RALL is children's phonological memory, defined as the capability to construct a phonological representation of speech sound sequences and to temporarily hold this representation active in memory for further processing (Gathercole & Baddeley, 1990; for a review on the relationship between phonological memory and word learning, see Gathercole, 2006). Phonological memory has been found to predict both L1 and L2 vocabulary learning (Baddeley, Gathercole, & Papagno, 1998; Gathercole, 2006; Gathercole & Baddeley, 1990; Masoura & Gathercole, 2005; Service, 1992; Verhagen, Boom et al., in press). Phonological memory may aid L2 vocabulary learning in particular if the learner is a novice and still has a limited L2 vocabulary (Cheung, 1996; Masoura & Gathercole, 2005). In this case, the learner cannot rely on vocabulary-related mechanisms of storing new phonological information (Metsala, 1999; Verhagen, Boom et al., in press) and cannot use semantic associations between existing concepts in L2 and the new L2 words (Masoura & Gathercole, 2005; Papagno, Valentine, & Baddeley, 1991). In the present study, involving young children with virtually no knowledge of the L2, we expected that individual differences in phonological memory would moderate the effect of RALL such that children with larger phonological memory capacity would benefit more from RALL than children with more limited phonological memory capacity.

Finally, language learning in both L1 and L2 may depend on general learning abilities, in particular selective attention seen as the core of executive functions and working memory, also referred to as learning-related skills (Cowan, 2014; McClelland, Cameron, Wanless, & Murray, 2007; Mulder, Hoofs, Verhagen, van der Veen, & Leseman, 2014). Selective attention, defined as a domain-general, effortful mechanism of perceptual focusing, may help to filter relevant from irrelevant information in the encoding stage of linguistic information processing. Attention may also be involved in selectively refreshing memory representations to prevent decay, keeping these representations available for integration in long term memory (Cowan, 2014). Although language learning is thought to depend in part on automatic implicit processes (e.g., statistical learning), studies with infants and older children have revealed that attention can strengthen implicit learning, for instance, by focusing perception on relevant cues such as the movements of a speaker's mouth to learn how to produce speech sounds (Lewkowicz & Hansen-Tift, 2011) or on the phonological distributional cues that signal word boundaries in the speech stream (Stevens & Bavelier, 2012). Selective attention has been related to incidental word learning in L2 (Godfroid, Boers, & Housen, 2013; Godfroid & Schmidtke, 2013). However, especially in L2 learning at a later age, as in sequential bilingualism, learning may also depend on explicit processes that require attention effort (e.g., the Noticing Hypothesis, Schmidt, 1990). Moreover, managing attention may be especially crucial in complex learning situations such as RALL. For example, in RALL, as in the current study as will be explained later, the setup often consists of a robot and a tablet as a mediating device that displays the educational content. In such a situation the learner must focus on the relevant information while ignoring or resisting distracting information, which involves selective attention (Robinson, 1995). In sum, attention skills may be required for a number of processes, implicit as well as explicit, that underlie word learning in L2. Differences between children in attention skills, therefore, can be hypothesized to moderate the effectiveness of learning interventions, including robot-supported L2 learning. With regard to the current study, we expected children high in selective attention to learn more from robot-assisted L2 learning than children who are low in this skill.

This Study

In the present study, we investigated (1) whether children benefited from the presence of a robot as tutor when learning L2 vocabulary, and (2) whether individual differences in prior L1 knowledge, phonological memory, and selective attention moderated the effects of the robot[9]. Native Dutch speaking children were taught L2 English vocabulary in the domains of mathematical and spatial language in a series of seven short, individually administered lessons. Children were taught the words through language games on a tablet in one of three conditions: (a) a tablet only; (b) a tablet and a robot that used deictic (pointing) gestures; or (c) a tablet and a robot that used both deictic and iconic gestures (gestures depicting the target word). In addition, (d) a control group of children was included who did not receive the vocabulary training but played dancing games with the robot instead. Children were recruited in the kindergarten departments of nine primary schools and within schools randomly assigned to one of the four conditions. The possible added value of iconic gestures in condition (c) was of particular interest as several studies have revealed the importance of iconic gesturing as support to L2 vocabulary learning (Macedonia et al., 2011; Rowe, Silverman, & Mullan, 2013; Tellier, 2008). Moreover, a recent study demonstrated that

[9] Parts of these data have been published in Vogt et al. (2019).

gestures also support L2 vocabulary learning in robot-assisted learning situations (de Wit et al., 2018). To assess the effects of the experimental conditions relative to the control condition, and to determine whether the three experimental conditions differed in effectivity, we tested children's knowledge of the English words immediately after the lessons series was completed and three to five weeks later. As possible moderators, we included measures of children's L1 vocabulary knowledge, phonological memory, and selective attention.

We hypothesized that children in the experimental conditions would learn more L2 words than children in the control condition, and that children in the robot-assisted conditions would learn more L2 words than children in the tablet-only condition. The robot's presence was expected to be motivating throughout the lesson series, leading to enhanced learning, and the robot's deictic and iconic gestures were expected to support word learning additionally, resulting in larger learning gains compared to the tablet-only condition. Also, based on the research into the role of gesturing, we expected that children would learn more L2 words in the condition in which the robot used iconic gestures than in the condition in which it did not.

As our study is, to the best of our knowledge, the first to investigate the moderating effects of children's language, memory and attention skills in RALL, we had only a general hypothesis. Based on findings in language learning research, discussed above, we expected that children with larger L1 vocabulary knowledge, larger phonological memory capacity, and a higher level of selective attention would learn more words across all experimental conditions than children scoring lower on these skills. Regarding possible differences in moderator effects between the three experimental conditions, we had no hypotheses beforehand, but considered a number of (contrasting) possibilities. Children low in language learning abilities (e.g., L1 vocabulary and phonological memory) may profit from more support and, therefore, show larger gains in the robot-assisted conditions than in the tablet-only condition. This may in particular hold for the robot with iconic-gestures condition where the robot provides additional non-verbal support. Children high in language learning abilities may not need extra support and, therefore, not show a difference in learning effects between the experimental conditions.

In contrast, children low in selective attention may have more difficulty to focus on the task demands the more complex the learning situations is, which would favor their word learning in the relatively simple tablet-only condition compared to the relatively complex robot conditions. For children high in selective attention, increased situational complexity, as in both robot conditions, may either not matter (no difference in the moderator effect between the experimental conditions) or even provide additional opportunities for learning (increased word learning in the robot-assisted conditions compared to the tablet-only condition because of the enriched information in these conditions).

To summarize, the current study addressed a number of questions and hypotheses by conducting a within-schools randomized controlled trial to compare native Dutch-speaking children's learning gains in English mathematical and spatial words across the different conditions. In addition, the study examined the possible moderating effects of individual differences in language, memory and attention skills on English word learning across the experimental conditions. By using a multiple-session design of word learning in the domains of mathematical and spatial language, we could not only control for a possible novelty effect, but also evaluate an educationally realistic robot-assisted vocabulary training program that is similar to non-robot vocabulary training provided in (pre)schools.

Method

Participants

One hundred and ninety-three monolingual Dutch preschoolers (95 girls) with an average age of 68.4 months (range 59 – 81 months, *SD* = 4.7 months) participated in the study. They were recruited from nine different schools in the Netherlands and were randomly assigned within schools to one of the four conditions, while ensuring a similar gender distribution over conditions. Table 1 displays the background characteristics of the children divided over the four conditions. There were no significant differences between conditions in parental education, age, and gender, all *p*s > .303. Eleven additional children started the lessons but did not complete them due to illness, technological problems, or because they did not want to participate anymore (*n* iconic-gesture condition = 6, *n* no-iconic-gesture condition = 3, *n* tablet-only condition = 2). Three additional children were pre-tested but excluded from the experiment because they knew more than half of the target words during the pre-test. One additional child had a different home language in addition to Dutch (German), and was excluded from analysis as well. Informed consent for all children was obtained from parents/caretakers prior to data collection.

Design

The experiment consisted of a pre-test, a lesson series, and two post-tests (see Figure 1 for an overview of the experiment). All children participated in a group introduction prior to the pre-test (see Procedure for more information). In the pre-test, we measured children's knowledge of the target words and their L1 vocabulary knowledge, phonological memory, and selective attention. The training was administered in one of four conditions, and had a between-subject design. In the experimental conditions, children played language games on a Microsoft Surface

Table 1. Background Characteristics of the Children in the Four Conditions

	Iconic gesture	No iconic gesture	Tablet-only	Control
N	54	54	53	32
n girls	22	26	29	18
M age (*SD*) in months	68.4 (4.8)	68.5 (4.7)	69.1 (4.4)	66.9 (4.7)
Age range in months	60-81	59-79	61-79	59-79
Parental education				
Academic level	74%	72%	60%	66%
Vocational level	20%	26%	33%	24%
Secondary school	6%	2%	7%	10%

Note. Information on parental education of both parents was gathered through a questionnaire with a response rate of 65.8%, thus for 127 out of 193 children (*n* iconic-gesture condition = 40, *n* no-iconic-gesture condition = 32, *n* tablet-only condition = 34, *n* control condition = 21).

tablet: (a) with a NAO robot that used iconic and deictic gestures (see the Robot section for more information on the robot used); (b) with a NAO robot that used only deictic gestures; or (c) by themselves. Children received usually two lessons per week and the training took on average 24 days (*SD* = 5.5 days). Children in a fourth, control condition did not play language games but danced with the robot instead during three sessions. Children's learning gains were measured in a game concluding each lesson (which will not be discussed further in this chapter), and in a post-test one or two days after the training and a second post-test two to four weeks after the first post-test (*M* = 18.9 days, *SD* = 3.6 days) to measure retention over a longer period.

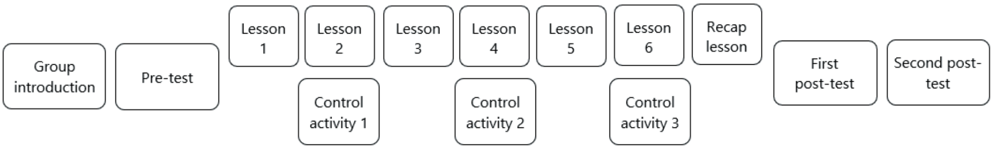


Figure 1. Overview of the experiment.

L2 Vocabulary Lessons

The lessons series consisted of seven individual lessons: six lessons in which new L2 vocabulary was provided, and one recap lesson in which all target words were repeated. Five or six target words were taught within each lesson, resulting in a total of 34 target words. The target words were chosen such that they were part of early mathematical and spatial language. The overall theme of the lesson series was an area to be explored, with different locations for each lesson. The locations were chosen such that they were familiar and relevant to young children. See Table 2 for an overview of the lesson series, the locations, and the target words.

Table 2. Overview of the Lesson Series and Target Words

Lesson	Location	Target words
One	Zoo	One, two, three, add, more, most
Two	Bakery	Four, five, take away, fewer, fewest
Three	Zoo	Big, small, heavy, light, high, low
Four	Fruit shop	On, above, below, next to, fall
Five	Forest	In front of, behind, walk, run, jump, fly
Six	Play ground	Left, right, catch, throw, slide, climb
Seven	Photo book	Repetition of all target words

Each lesson consisted of three parts. First, the child was greeted, a reference was made to the previous lesson, and the location of the current lesson was introduced. Then, the new target words were modelled. New target words were first introduced by a pre-recorded speech sample of a native (Canadian) English speaker. The child was asked to repeat the target word, as this benefits productive recall of target words (Ellis & Beaton, 1993). Then, the child had to perform several tasks on the tablet to practice the target words, for example, putting two elephants in a cage to practice the word “two”. The tasks to practice the target words differed per target word. Some target words required manipulations on the tablet, while others allowed for more physical activity. For example, children were asked to act out running when learning the word “running”. The lesson concluded with a short test, to measure immediate learning gains. We will not discuss these immediate tests in this chapter.

Each target word was repeated ten times throughout the lesson: nine times by the robot, and once by the native English speaker when it was introduced. Each target word reoccurred once in the following lesson and twice in the recap lesson. During the recap lesson, a photo book appeared on the tablet, which showed print screens from the previous lessons. Children had to practice repeating the target words once more during this lesson.

Robot

The robot used was a NAO robot. The NAO robot is a 58cm tall humanoid robot. The robot was sitting in crouch position during the lesson series in a 90 degree angle to the right of the child, which was sat on the floor facing the tablet that was positioned on an elevated surface.

The robot’s responses had been preprogrammed, such that its responses and behaviors were consistent for all children. The robot behaved almost fully autonomously. The only function that was controlled by the experimenter was voice detection, as automatic speech recognition systems do not work reliably for children (Kennedy et al., 2017). The experimenter indicated, using a graphical user interface on a laptop computer, whether the child had repeated the target words or not when asked to do so. The laptop computer was not in direct sight of the child (see Figure 2). The robot was introduced as Robin (which is a unisex name in Dutch), being a peer who was going to learn English words together with the children.

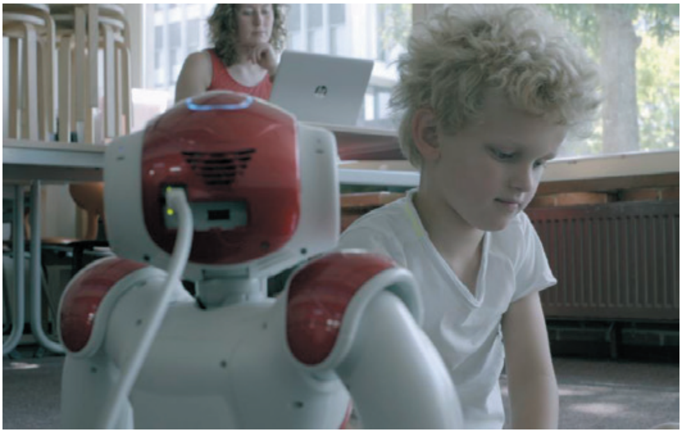


Figure 2. A child playing with the robot.

The robot acted as a slightly more knowledgeable peer who understood the game usually faster than the child did. As such, the robot performed several behaviors during the training: 1) talking to the child and explaining the tasks of the lesson; 2) pronouncing the target words; 3) providing feedback on the actions of the child; 4) pointing to the tablet while explaining what to do; 5) performing required manipulations in case the child failed to perform a specific task. In case of the latter, the robot moved its arm above the tablet and any required manipulations 'magically' occurred.

To ensure that the content and structure of the lessons were the same between the different conditions, in the tablet-only condition, the robot's voice was redirected through the tablet's speakers, and the robot itself was hidden from sight. Thus, the robot was used ‘behind the scenes’ to operate the system, but children only saw and interacted with the tablet. In the robot-assisted conditions children thus interacted with the robot and the tablet, whereas in the tablet-only condition they interacted with the tablet only.

Measures

Pre-test translation task

To measure whether children knew the L2 English target words prior to the lesson series, we administered a translation task. In this task, children heard the 34 English target words one by one and were asked to translate them to Dutch. The target words were prerecorded by a native speaker and played through a laptop computer. Two versions of this task were used, differing in word order. The first list of words was created by listing the target words randomly, and a second list was created by reversing the first list. Children were awarded one point per correct answer, yielding a total of 34 points. Cronbach’s alpha showed that the reliability for the task was excellent, α = .96.

Post-test translation tasks

To measure how many L2 English target words the children learned during the lesson series, we administered two translation tasks: one from English to Dutch and one from Dutch to English. The task was the same as the pre-test translation task, except that children now also had to translate the words from Dutch to English. Both tasks were administered twice, once during the first post-test and once during the second post-test. Children were awarded one point per correct answer, resulting in a total of 34 points per task. Cronbach’s alpha showed that the reliability for both tasks was excellent, α = .94 for the first post-test andα = .95 for the second post-test for the

English-to-Dutch translation task, and α = .97 for the first post-test and α = .98 for the second post-test for the Dutch-to-English translation task.

Post-test comprehension task

We administered a comprehension task to measure children's receptive knowledge of the target words taught. The comprehension task was a picture-selection task in which we presented children with three images (still photos for most words, or short films in the case of verbs) on a laptop screen. Children then had to select the image corresponding to the target word they heard. Again, pre-recorded speech was used. A bilingual native English-Dutch speaker pronounced a Dutch carrier sentence *"waar zie je"* ("where do you see") followed by the target word in English. There were three trials for each target word, with different distractors each time. We selected half of the target words for this task to reduce children's fatigue, as a comprehension task consisting of all items would have been too long for the children. The target words included in the tests were chosen such that words from each lesson were included and that different types of words (verbs, adjectives, prepositions) were included. Two versions of this task were used, differing in word order: The first list of words was created by listing the target words randomly, and a second list was created by reversing the first list. Children were awarded one point per correct answer, resulting in a maximum score of 54 points. This task was administered during the first and second post-test. Cronbach's alpha showed that the reliability of the task was good, α = .84 for the first post-test and α = .87 for the second post-test.

L1 vocabulary

We used the Dutch version of the Peabody Picture Vocabulary Test (PPVT) to measure children's Dutch receptive vocabulary knowledge (Dunn, Dunn, & Schlichting, 2005). This task is a picture-selection task in which children are presented with four pictures and are asked to select the picture corresponding to a word said by the experimenter. The task contains a total of seventeen sets, of which each set consists of twelve items. The test is adaptive, such that the starting set is chosen depending on the age of the child, and is stopped when the child makes nine or more errors within one set. The L1-vocabulary test is age-normed, with a mean of 100 and a standard deviation of 15. Cronbach's alpha is described in the test manual to be between .92 and .94. We used standardized scores in our analyses.

Phonological memory

The Quasi-Universal Nonword Repetition Task (Q-U NWRT) was used to measure phonological memory (Boerma et al., 2014; Chiat, 2015). The Q-U NWRT is a computerized task appropriate for young children, consisting of twelve items. Children hear a previously recorded, non-existing word via a laptop computer, and are asked to repeat it. Children receive two practice items (two one-syllable nonwords) before starting. Children's responses were scored online by the experimenter and received one point for each word that they repeated correctly, yielding a maximum score of twelve. Cronbach's alpha showed that the reliability of this task was satisfactory, α = .76. Ten percent of the data was scored by an additional researcher based on video recordings of the test. Inter-rater reliability was good with 89% agreement, κ = .74 (95% CI, .663 to .819), p < .001.

Selective attention

A computerized visual search task was used to measure selective attention (Mulder et al., 2014). In this task, children were shown a display of animals on a laptop screen consisting of elephants, bears, and donkeys that were similar in color and size. Children were asked to find as many elephants as possible among distractor animals. Children were given three practice items and four test items that increased in difficulty. In the first two test items, 48 animals appeared on a six by eight grid. In the third item, 72 animals (similar in size to the first two test items) appeared on a nine by eight grid. In the last item, 204 animals (smaller in size than in the other three test items) appeared on a 12 by 17 grid. There were eight targets (elephants) in total in each test item. Each test item lasted 40 seconds. The experimenter encouraged children to search as quickly as possible and gave feedback according to a protocol. Elephants that were found were crossed off with a line by the experimenter. The number of targets located correctly per round were calculated and averaged across items, resulting in a maximum score of eight. Cronbach’s alpha showed that the reliability of this task was good, α = .86.

Procedure

Group introduction of robot

Prior to any individual sessions, the robot was introduced to all children in a group session. The robot introduced itself and did a dance with the children. The group introduction served to familiarize children with the robot, and reduce potential anxiety during the individual sessions.

Pre-test

All children were tested individually by a trained experimenter in a quiet room in their schools. Children were administered the tasks in the following order: PPVT, pre-test translation task, selective-attention task, and quasi-universal non-word repetition task. Furthermore, a perception questionnaire was administered (also during the first post-test) which measured the degree to which children anthropomorphized the robot. This questionnaire did not measure language skills or learning gains. The results of this questionnaire are beyond the scope of the current chapter but will be discussed in

Chapter 6. The pre-test session lasted 30-40 minutes. Children got a sticker in reward for each task.

L2 vocabulary lessons

Each lesson was administered individually in a quiet room in the children's schools. At the start of the first session, the experimenter explained how the child could perform the requested actions on the tablet during the lessons (e.g., swiping and tapping), and helped the child to play the game. The experimenter was always present during the lessons to help children if needed, and to control the robot. The lesson could be paused if children needed a break. Each lesson lasted 15-20 minutes.

Control activities

Children in the control condition participated in a total of three activities with the robot, each administered individually in a quiet room in the children's schools. In each session, the robot greeted the children, did a dance together with the child, and said goodbye. Each session lasted around five to ten minutes.

First and second post-test

Children were administered the various tasks in the following order: the English-to-Dutch translation task, the Dutch-to-English translation task, and the comprehension task. During the first post-test the anthropomorphism questionnaire, further discussed in Chapter 6, was also administered. Each session lasted around 30 minutes. Children got a sticker in reward for each task completed.

Analyses

We ran a MANOVA to compare the four groups of children on L1 vocabulary knowledge, phonological memory, selective attention, and pre-test scores. Children's scores on the comprehension task were compared against chance level (33%). To investigate differences in learning gains between the three conditions, we ran mixed-effect logistic regression models in the statistical package R (R Core team, 2017) using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015). Dependent variables were children's binary (correct/incorrect) scores on the translation tasks and the comprehension task. The analyses were run separately for the translation tasks and the comprehension task, as they were assumed to measure different types of vocabulary knowledge. For both tasks, both assessments (the first and second post-test) were included.

Random factors and slopes were included through the method of model comparisons, in which the most parsimonious model that best fitted the data was identified. In order to compare models, likelihood ratio tests were performed that compared the goodness of fit using the ANOVA function in the base package (R Core Team, 2017). In this way, the final model was selected by checking whether the *p*-value from the likelihood ratio test was significant. For the translation tasks, 'subjects', 'target words', and 'test item number' were included as random factors, and random slopes for target words (condition*target word). For the comprehension task, 'subjects', 'target words', and 'test item number' were included as random factors, and no random slopes were included as models including random slopes did not converge.

In all models, orthogonal sum-to-zero contrast coding was applied to our categorical fixed effects (i.e., condition, post-test, language), and all continuous variables (i.e., vocabulary knowledge, phonological memory, selective attention) were centered around zero (Baguley, 2012, pp. 590 – 621). For time, the first post-test (coded as -0.5) was contrasted with the second post-test (coded as 0.5). For condition, there were three contrasts: Contrast 1 contrasted the three experimental conditions (each coded as 0.25) with the control condition (coded as -0.75); Contrast 2 contrasted the two robot-assisted condition (each coded as -0.33) with the tablet-only condition (coded as 0.66); and Contrast 3 contrasted the iconic-gesture condition (coded as -0.5) with the no-iconic-gesture condition (coded as 0.5).

For both models, the factors condition (control, tablet-only, robot without iconic gestures, and robot with iconic gestures) and time (first and second post-test) were included as fixed effect factors, with an interaction between them. For the translation task, target language (from English as source to Dutch as target, and vice versa) was included as an additional factor. The models were re-run separately for each of the three moderator variables (L1 vocabulary knowledge, phonological memory, and selective attention). They were included as a fixed effect factor in interaction with condition. The full results of each model can be found in the Appendix. To reduce the risk of Type-1 error when conducting multiple comparisons, we present and discuss all results with *p*-values of $p < .05$ for the models including the moderator variables, but only interpret these results further in the Discussion section when $p < .01$.

Results

Descriptive Analyses

Table 3 displays the descriptive statistics for all the variables included in the analyses for the children in each condition separately. A MANOVA showed no differences in L1 vocabulary, phonological memory, selective attention, and English vocabulary pre-test scores between the conditions, $F(12, 397) = 1.75$, $p = .054$, $\eta_p^2 = .05$. For all conditions, children scored above chance level on the comprehension task on both the first and second post-test, all *p*s < .001, range *d*s = 1.49-2.83.

Table 3. Mean Outcomes (Standard Deviation) of the Children in the Four Conditions

		Iconic	No iconic	Tablet-only	Control
Pre-test	L1 vocabulary	108.13 (12.54)	108.67 (11.83)	105.77 (11.92)	108.88 (13.96)
	Phonological memory	10.08 (2.97)	11.33 (2.86)	11.08 (2.13)	10.16 (3.22)
	Selective attention	6.48 (0.65)	6.82 (0.58)	6.67 (0.64)	6.61 (0.82)
	Translation En-Du	3.41 (3.05)	3.59 (3.14)	3.98 (2.74)	2.81 (2.83)
First post-test	Translation En-Du	7.54 (5.14)	7.83 (4.94)	7.91 (4.63)	3.81 (3.21)
	Translation Du-En	6.09 (4.15)	6.54 (4.28)	6.64 (4.01)	3.16 (2.27)
	Comprehension	29.39 (5.78)	29.50 (6.13)	29.53 (6.40)	25.03 (6.66)
Second post-test	Translation En-Du	8.20 (4.98)	8.02 (4.92)	8.57 (4.61)	4.34 (3.22)
	Translation Du-En	6.57 (4.60)	6.44 (4.59)	6.75 (4.22)	3.47 (2.13)
	Comprehension	30.54 (6.26)	29.69 (6.61)	30.30 (6.55)	26.00 (6.04)

Note. The L1-vocabulary test is age-normed, with a mean of 100 and a standard deviation of 15. The maximum scores were 16 for the phonological-memory test, 8 for the selective-attention test, 34 for each translation task, and 54 for the comprehension task (chance level for the latter task was 18).

Learning Gains and Effects of Robot Presence

Table 4 shows the results of the main models without moderator variables. We found a main effect of condition, with children in all three experimental conditions outperforming children in the control condition on both the translation tasks and the comprehension task. Thus, children learned from the vocabulary training. There were no significant differences between the three experimental conditions, however. Children did not learn more target words in the robot-assisted conditions as compared to the tablet-only condition. There also was no additional benefit of the robot making

Table 4. Results from the Mixed-Effects Logistic Regression Model with Accuracy Scores from the Translation Tasks or the Comprehension Task as the Dependent Variable, Condition as Between-Subjects Fixed Effect, and Time (and Target Language for Model Including the Translation Tasks) as Within-Subjects Fixed Effect

	Translation tasks				Comprehension task			
	ß	SE	*z*	*p*	*ß*	SE	*z*	*p*
(Intercept)	-3.19	0.43	-7.46	< .001	0.25	0.22	1.10	.270
Condition c1	1.90	0.38	4.97	< .001	0.35	0.12	3.07	.002
Condition c2	0.12	0.29	0.41	.681	0.02	0.11	0.14	.890
Condition c3	-0.05	0.33	-0.15	.879	-0.10	0.17	-0.55	.580
Post-test	0.16	0.05	3.28	.001	0.05	0.04	1.09	.276
Language	-0.56	0.05	-11.11	< .001	NA	NA	NA	NA
Condition c1 * time	-0.15	0.15	-1.03	.301	-0.06	0.09	-0.60	.552
Condition c2 * time	0.05	0.10	0.47	.639	0.04	0.09	0.44	.660
Condition c3 * time	-0.20	0.12	-1.70	.089	-0.01	0.14	-0.06	.954

Note. Condition c1: contrast experimental vs. control. Condition c2: contrast robot-assisted vs. tablet-only. Condition c3: contrast with vs. without iconic gestures.

iconic gestures. Furthermore, a main effect of time of testing was found, with children obtaining higher scores on the second post-test than on the first post-test. Note, however, that this effect was only found for the translation tasks, and not for the comprehension task. Last, a main effect of language was found for the translation tasks, with children obtaining higher scores on the English-to-Dutch translation task than on the Dutch-to-English translation task.

Effects of Language and Attention Skills

In this section, we report and interpret effects that were significant at $p < .05$. In the next section, we will only integrate findings in our general discussion that were significant at $p < .01$. All outcomes can be found in the Appendix. First, we discuss the moderator effects of L1 vocabulary, phonological memory, and selective attention on the comparison of the experimental conditions versus control condition. Then, we will discuss the moderator effects on the comparisons of the robot-assisted versus tablet-only conditions, and on the iconic-gestures versus no-iconic-gestures conditions.

Regarding the experimental versus control conditions, the models of the translation tasks showed statistically significant interactions between the moderator variables and condition. There were positive main effects for L1 vocabulary, $ß$ =72.47, SE = 8.99, z = 8.06, p < .001, phonological memory, $ß$ = 22.13, SE = 8.07, z = 2.74, p = .006, and selective attention, $ß$ = 36.46, SE = 6.47, z = 5.63, p < .001, but only for children in the experimental conditions, and not for those in the control condition. This interaction was to be expected, as only children in the experimental conditions received an L2 vocabulary training in which they could benefit from these skills. Note that the effects were only found for the translation tasks and not for the comprehension task.

Regarding the robot-assisted versus tablet-only conditions, differences in translation task scores were found between the robot-assisted and tablet-only conditions for two of the three moderators, that is, L1 vocabulary, *ß* = -24.52, *SE* = 10.94, *z* = -2.24, *p* = .025, and phonological memory, *ß* = 26.72, *SE* = 10.53, *z* = 2.54, *p* = .011. Children with larger L1 vocabularies learned more words in the robot-assisted conditions than in the tablet-only condition, while this effect was opposite for phonological memory: Children with better phonological memory learned more in the tablet-only condition than in the robot-assisted conditions.

Last, moderator effects were found for both robot-assisted conditions. Children with larger L1 vocabularies learned more target words in the condition in which the robot did not use iconic gestures than in the condition in which it did, as was indicated by the models run on the translation tasks, *ß* = 31.99, *SE* = 9.16, *z* = 3.49, *p* < .001, and comprehension task, *ß* = 19.56, *SE* = 6.43, *z* = 3.05, *p* = .002. The same pattern was found for phonological memory, for both the translation tasks, *ß* = 40.52, *SE* = 14.99, *z* = 2.70, *p* = .007, and the comprehension task, *ß* = 12.85, *SE* = 5.67, *z* = 2.27, *p* = .023. Selective attention showed an opposite pattern: Children with better selective attention showed higher performance in the condition in which the robot used iconic gestures than in the condition in which it did not, but only on the translation tasks, *ß* = -41.25, *SE* = 8.75, *z* = -4.71, *p* < .001.

In summary, various moderator effects were found for the language learning related individual child characteristics included in this study. Note, however, that some of these effects were relatively small and not significant at *p* < .01, and, therefore, will not be further interpreted. The main findings are as follows. In addition to the main effect of condition on children's English vocabulary knowledge, there were extra benefits associated with children having a larger L1 vocabulary, a larger phonological memory capacity and a higher level of selective attention in the experimental conditions as opposed to the control condition. Differential effects were found for each of the moderator variables regarding the three experimental conditions. Children with larger L1 vocabularies and better phonological memory learned more when the robot did not use iconic gestures than when it did use iconic gestures. Children with better selective attention showed higher learning gains when the robot used iconic gestures than when it did not use iconic gestures.

Discussion

The aim of our experiment was to investigate (1) whether a social robot supports preschool children’s L2 vocabulary learning, and (2) whether this effect is moderated by individual differences in relevant language learning skills. Whereas the majority of RALL studies report on single session studies with only a few target words (see Chapter 2), our study, using a social NAO robot, examined the potential of RALL to teach a more comprehensive vocabulary training program with a substantial number of target words (i.e., 34), to better match to the needs of (pre)schools. The multiple sessions allowed us also to control for the so called novelty effect, the often observed initially high but soon declining motivation of children to interact with a robot tutor, which can lead to an overestimation of the effectivity of robot-assisted learning. Children in the present study were taught L2 English vocabulary through seven lessons in the form of tablet games, which they played either: (a) alone; (b) together with a robot that used deictic gestures; or (c) together with a robot that used both deictic and iconic gestures. Furthermore, the children in the experimental conditions were compared to (d) a control group of children who did not play language games but played dancing games with the robot instead. We used random assignment within schools to place children in one of the four conditions, while ensuring a roughly equal distribution of gender over the conditions. The tablet was an essential device in the robot conditions as technical limitations, in particular the lack of accurate speech perception (Kennedy et al., 2017) and object recognition for the type of robot we used (Wallbridge et al., 2017) required this extra device to enable interaction and communication. The tablet-only condition, therefore, was meant to disentangle the effects of the robot from the effects of the tablet. To identify whether robots as L2 tutors benefitted all children equally or subgroups of children more, we included children's L1 vocabulary knowledge, phonological memory, and selective attention as potential moderators.

The first aim of the study was to determine whether robot assisted L2 learning was effective compared to a control condition without an L2 learning intervention and a condition with a language learning intervention but using only a tablet. We found that children learned significantly more in the three experimental conditions than in the control condition, as was expected. Thus, children indeed learned vocabulary from our lessons. Unexpectedly, however, there were no differences between the experimental conditions. Children in the robot-assisted condition did not learn more than children who worked with the tablet only. One previous RALL study that compared a vocabulary training with a robot to a training with a tablet found similar results (Kory Westlund et al., 2015). Similarly, a one-session study in which children were learning in a sorting task assisted by either a physical robot or a virtual robot found no advantage on task scores of the physically present robot either (Kennedy, Baxter, & Belpaeme, 2015a).

To be able to compare the robot-assisted and tablet-only conditions, the conditions were kept as similar as possible. The children in the tablet-only condition received identical verbal support, explanations, and feedback, voiced by the hidden robot via the tablet's speakers, as the children in the robot-assisted conditions. Therefore, the only difference was that children did not receive non-verbal support in the tablet-only condition through the robot's social presence and its (iconic and) deictic gestures, which we hypothesized beforehand to be potential advantages of the robot. This setup and in particular the prominent role of the tablet may explain the lack of additional benefits of the robot. As the content of the lesson series was displayed on the tablet and most actions had to be performed on the tablet screen, children focused mainly on the tablet and may not have sufficiently attended to the robot to benefit from its presence.

We expected the robot's presence to be motivating and engaging, but an additional potential explanation for the lack of added value may be that the robot, contrary to this expectation, distracted children from focusing on the educational content presented on the tablet rather than supporting engagement in the learning task. In previous research, robots have been found to be distracting if they displayed too much social behavior (Kennedy et al., 2015). The robot may have distracted children in our study even more due to the complexity of the tutoring situation and the time it took to perform deictic and iconic gestures, an issue to which we will return below. Note that for technical reasons we could not include a robot condition in which no tablet was used. As a consequence, we cannot draw firm conclusions regarding the effectivity of a robot tutor without tablet. Future studies, with improved technology that allows for a stand-alone robot condition are needed to be able to draw more definitive conclusions about the potential of robots as L2 tutors.

Furthermore, the expected benefit of a robot using iconic gestures to support children's L2 learning could not be confirmed. This contrasts with the findings in previous research showing the critical importance of gesturing in beginning language learners involving human tutors (Rowe et al., 2013) and also with the findings of a recent single-session RALL study (de Wit et al., 2018). A possible explanation is that the iconic gestures in our study were ambiguous, which was partly due to the more abstract nature of some of the target words, but also to the physical and technical limitations of the NAO robot. Gestures in the single-session study that taught L2 nouns (de Wit et al., 2018) may have been better perceivable and interpretable, than the gestures in our study that had to exemplify mathematical and spatial concepts. For example, it can be easier to have a robot depict a monkey (e.g., by making it scratch its head and armpit) than a count word using finger counting, because a NAO robot can only move all of its three fingers at once and, therefore, cannot count on its fingers similar to humans. Future work should use improved gesturing to examine the theoretically presupposed advantages of robots as embodied in the world over other technology. One possible avenue is to investigate whether children's re-enactment of the robot's gestures will benefit (L2) learning more, as a recent study suggests (Macedonia & von Kriegstein, 2012).

The second aim of the present study was to examine possible moderator effects that could indicate that some children profit more from RALL than other children, as has been suggested (Kanda et al., 2004). To this end, we included individual child characteristics deemed relevant for (second) language learning in our analyses: L1 vocabulary, phonological memory and selective attention. Several statistically significant moderator effects were found, both expected and unexpected. Below, to avoid capitalization on chance given the multiple comparisons, we only review the moderator effects that were significant at p < .01.

Regarding the overall effectiveness of the experimental conditions involving word learning lessons compared to the control condition without word learning, we found the expected moderator effects: Children scoring high on L1 language knowledge, phonological memory or selective attention, as assessed prior to the experiment, learned more from the experimental vocabulary lessons than children scoring low on these skills, in line with a vast literature that showed similar advantages in (second) language learning in general (Baddeley, Gathercole, & Papagno, 1998; Gathercole, 2006; Gathercole & Baddeley, 1990; Masoura & Gathercole, 2005;

Robinson, 1995; Schmidt, 1990; Service, 1992; Verhagen, Boom et al., in press; Wolter, 2006).

Regarding possible differences in moderator effects between the three experimental conditions, we had no hypotheses in advance but we considered a number of possible mechanisms. Significant moderator effects at $p < .01$ were found for the comparison of the robot with and without iconic gestures, in particular regarding the two translations tasks. Children with larger L1 vocabularies or larger phonological memory capacity as assessed prior to the experiment learned more English words than children with smaller L1 vocabularies or less phonological memory capacity in the robot condition without iconic gestures compared to the robot condition with iconic gestures. Note that these moderator effects were observed in addition to positive main effects of both conditions compared to the control condition. A possible explanation is that the robot condition without iconic gestures provided less support to the language learning and that the children, as a consequence, had to rely more on their own, but varying, language learning abilities in this condition than in the robot condition with extra support in the form of iconic gestures. The extra support through iconic gesturing, moreover, decreased the overall speed of the interaction because performing the gestures took extra time per target word, and thus led to more time to process the presented information, which may have leveled-off the effect of individual differences in language learning abilities. If true, the findings suggest that iconic gesturing in RALL may support children with weaker language learning abilities.

Selective attention, with a significant moderator effect at $p < .01$, showed an opposite pattern. Children high in selective attention learned more English words than children low in selective attention in the robot condition with iconic gestures compared to the robot condition without iconic gestures. Note again that the moderator effect was found in addition to positive main effects of both experimental conditions relative to the control condition. A possible explanation points to the previously discussed ambiguity of the iconic gestures and the distracting effect this may have had on children's word learning in this study (cf. Robinson, 1995; Schmidt, 1990). Children high in selective attention may have been better able to profit from the additional cues, which assumingly required attention effort to perceive and interpret, and/or may have been less distracted by the extra information provided. Children low in selective attention may have been less capable in figuring out what the meaning was of the gestures and/or were more easily distracted by the gestures and the extra time it took to perform these gestures. If true, this suggests that implementing iconic gestures benefits children with good attention skills, but disadvantages children with less good attention skills. Whether this differential effect pertains to the difficulties with the perception and interpretation of the (ambiguous) iconic cues or to the distracting effect of the extra time these gestures took, or to both, cannot be decided based on the current data, but the findings, in general, suggest that both improving the quality of the gestures and limiting the time it takes to perform them, could improve the effectivity of RALL with iconic gesturing, at least with regard to individual differences between children in selective attention.

The two explanations offered for the opposite pattern of moderator effects we found are seemingly contradictory. For example, more information and time as consequence of performing iconic gestures may be beneficial for children with smaller vocabularies and less phonological memory capacity, but disadvantageous for children with low attention skill. Although language learning abilities, as exemplified here by L1 vocabulary knowledge and phonological memory, correlate with selective attention, they constitute distinct types of skills, leading in combination to different language learner profiles. Designing RALL situations to fit optimally to individual learner profiles could be a next step in improving the effectivity of RALL.

Finally, no moderator effects were found in the control condition, which was expected because the control condition did not involve a word learning intervention. The control condition, however, did involve an immediate and delayed post-test, similar to the experimental conditions. The lack of moderator effects in the control condition, therefore, supports the interpretation of the moderator effects in the experimental conditions as pertaining to the learning process, not to the test taking.

Limitations to the robot intervention

The overall results of the present study reveal that using robot tutoring in L2 learning programs for young children has still a long way to go. The learning gains were, overall, modest (on average 8 of 34 words were learned after seven lessons of 15 to 20 minutes, with frequently repeated exposure of the target words and a recap session at the end of the lesson series) and there was no overall added value of using a robot. The learning gains were much smaller than have been found in traditional vocabulary training interventions involving human tutors (Marulis & Neuman, 2010). There are several possible explanations. Designing the lesson series around the NAO robot, given the current state of technology, put severe constraints on the design of the lessons, required the use of a tablet for communication, and necessitated strong standardization. Traditional vocabulary training interventions may include more diverse activities that benefit learning and motivation, such as moving around and joint playing with objects. A particularly severe limitation was that the robot was not capable of understanding children and to follow-up on their meaningful, but sometimes idiosyncratic and unpredictable ways of responding, nor to truly adapt its responses and feedback to the children beyond mere encouragement or, in case of errors, mere repetition of the instructions. In contrast, a human tutor, in particular a competent, child-centered teacher, monitors children's comprehension and constantly adapts instruction, guidance and feedback to the understanding and support needs of the children. They may provide additional explanations when needed, quickly proceed to new materials if children show understanding of the previous materials, and keep children engaged during the learning task. The robot in our study could not monitor children's comprehension and engagement during the task. Children in our study all received the same instructional input, except for additional but standardized feedback if they did not manage to complete a task.

Conclusion

Our study is one of the first to investigate a robot's added value to L2 vocabulary learning in a multiple sessions and well-powered experiment. The lack of benefits of the robot suggests that the most important advantages of robots over other forms of technology, that is, the possibilities of play and interaction due to their physical presence, mostly exist in theory and cannot be effectively implemented in practice yet. Moreover, our study is the first to investigate individual differences in children's language and attention skills as moderators of a robot's beneficial effects. Taken together, the results suggest that the study of individual differences and moderators is highly relevant. It is likely that the effects of the robot are different for different children and adaptation to children's learning profiles is warranted. Indeed, one of the real advantages of robots is that they can play different roles for different types of learners if programmed to do so.

The present results should be replicated before any firm conclusions can be drawn. Future studies should include individual child factors and look for differential effects of technology. The study of individual differences is standard practice in educational sciences and developmental psychology, and could add to studies on the design of adaptive robots for educational practice.

Appendix

Below, the tables for each of the models and each of the language and attention skills (i.e., L1 vocabulary, phonological memory, and selective attention) can be found. The "*ß*" is an indicator of the effect size.

Results for L1 Vocabulary

Table A.1. Results from the Mixed-Effects Logistic Regression Model with Accuracy Scores from the English-Dutch Translation Task and Dutch-English Translation Task as Dependent Variables, Condition * L1 Vocabulary as Between-Subjects Fixed Effects, and Time and Language as Within-Subjects Fixed Effects

	ß	SE	*z*	*p*
(Intercept)	-3.19	0.43	-7.47	< .001
Condition contrast 1	1.86	0.37	4.99	< .001
Condition contrast 2	0.20	0.28	0.70	.483
Condition contrast 3	-0.10	0.32	-0.31	.757
Time	0.16	0.05	3.28	.001
Language	-0.56	0.05	-11.11	< .001
L1 vocabulary	25.19	8.96	2.81	.005
Condition contrast 1 * time	-0.15	0.15	-1.03	.302
Condition contrast 2 * time	0.05	0.10	0.47	.640
Condition contrast 3 * time	-0.20	0.12	-1.70	.090
Condition contrast 1 * L1 vocabulary	72.47	8.99	8.06	< .001
Condition contrast 2 * L1 vocabulary	-24.52	10.94	-2.24	.025
Condition contrast 3 * L1 vocabulary	31.99	9.16	3.49	< .001

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures.

Table A.2. Results from the Mixed-Effects Logistic Regression Model with Accuracy Scores from the Comprehension Task as a Dependent Variable, Condition * L1 Vocabulary as Between-Subjects Fixed Effects, and Time as a Within-Subjects Fixed Effect

	ß	SE	*z*	*p*
(Intercept)	0.24	0.22	1.09	.275
Condition contrast 1	0.34	0.11	3.05	.002
Condition contrast 2	0.02	0.11	0.22	.829
Condition contrast 3	-0.11	0.17	-0.66	.513
Time	0.05	0.04	1.08	.278
L1 vocabulary	4.01	4.13	0.97	.332
Condition contrast 1 * time	-0.06	0.09	-0.60	.550
Condition contrast 2 * time	0.04	0.09	0.44	.657
Condition contrast 3 * time	-0.01	0.14	-0.05	.958
Condition contrast 1 * L1 vocabulary	11.01	6.59	1.67	.095
Condition contrast 2 * L1 vocabulary	-9.86	6.67	-1.48	.140
Condition contrast 3 * L1 vocabulary	19.56	6.43	3.05	.002

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures.

Results for Phonological Memory

Table A.3. Results from the Mixed-Effects Logistic Regression Model with Accuracy Scores from the English-Dutch Translation Task and Dutch-English Translation Task as Dependent Variables, Condition * Phonological Memory as Between-Subjects Fixed Effects, and Time and Language as Within-Subjects Fixed Effects

	ß	SE	*z*	*p*
(Intercept)	-3.23	0.43	-7.57	< .001
Condition contrast 1	1.82	0.38	4.80	< .001
Condition contrast 2	0.12	0.29	0.40	.688
Condition contrast 3	-0.10	0.33	-0.30	.762
Time	0.16	0.05	3.18	.001
Language	-0.55	0.05	-10.93	< .001
Phonological memory	20.34	7.04	2.89	.004
Condition contrast 1 * time	-0.16	0.15	-1.08	.282
Condition contrast 2 * time	0.06	0.10	-0.56	.575
Condition contrast 3 * time	-0.18	0.12	-1.52	.129
Condition contrast 1 * phonological memory	22.13	8.07	2.74	.006
Condition contrast 2 * phonological memory	26.72	10.53	2.54	.011
Condition contrast 3 * phonological memory	40.52	14.99	2.70	.007

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures.

Table A.4. Results from the Mixed-Effects Logistic Regression Model with Accuracy Scores from the Comprehension Task as a Dependent Variable, Condition * Phonological Memory as Between-Subjects Fixed Effects, and Time as a Within-Subjects Fixed Effect

	ß	SE	*z*	*p*
(Intercept)	0.24	0.23	1.05	.294
Condition contrast 1	0.34	0.12	2.97	.003
Condition contrast 2	0.04	0.12	0.38	.708
Condition contrast 3	-0.09	0.18	-0.49	.621
Time	0.05	0.04	1.09	.277
Phonological memory	0.32	4.13	0.08	.939
Condition contrast 1 * time	-0.06	0.09	-0.60	.552
Condition contrast 2 * time	0.04	0.09	0.44	.659
Condition contrast 3 * time	-0.01	0.14	-0.06	.956
Condition contrast 1 * phonological memory	3.21	5.89	0.55	.586
Condition contrast 2 * phonological memory	-7.45	6.12	-1.22	.224
Condition contrast 3 * phonological memory	12.85	5.67	2.27	.023

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures.

Results for Selective Attention

Table A.5. Results from the Mixed-Effects Logistic Regression Model with Accuracy Scores from the English-Dutch Translation Task and Dutch-English Translation Task as Dependent Variables, Condition * Selective Attention as Between-Subjects Fixed Effects, and Time and Language as Within-Subjects Fixed Effects

	ß	SE	*z*	*p*
(Intercept)	-3.19	0.43	-7.47	< .001
Condition contrast 1	1.90	0.38	5.04	< .001
Condition contrast 2	0.13	0.29	0.46	.649
Condition contrast 3	-0.09	0.33	-0.26	.793
Time	0.16	0.05	3.28	.001
Language	-0.56	0.05	-11.11	< .001
Selective attention	20.58	6.66	3.09	.002
Condition contrast 1 * time	-0.15	0.15	-1.03	.302
Condition contrast 2 * time	0.05	0.10	0.47	.640
Condition contrast 3 * time	-0.20	0.12	-1.70	.089
Condition contrast 1 * selective attention	36.46	6.47	5.63	< .001
Condition contrast 2 * selective attention	-13.53	10.42	-1.30	.194
Condition contrast 3 * selective attention	-41.25	8.75	-4.71	< .001

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures.

Table A.6. Results from the Mixed-Effects Logistic Regression Model with Accuracy Scores from the Comprehension Task as a Dependent Variable, Condition * Selective Attention as Between-Subjects Fixed Effects, and Time as a Within-Subjects Fixed Effect

	ß	SE	*z*	*p*
(Intercept)	0.28	0.23	1.26	.210
Condition contrast 1	0.30	0.13	2.30	.021
Condition contrast 2	0.02	0.13	0.17	.862
Condition contrast 3	-0.21	0.19	-1.10	.272
Time	0.03	0.05	0.59	.557
Selective attention	9.06	4.61	1.96	.050
Condition contrast 1 * time	-0.01	0.12	-0.07	.946
Condition contrast 2 * time	0.05	0.10	0.47	.638
Condition contrast 3 * time	-0.02	0.15	-0.12	.908
Condition contrast 1 * selective attention	10.40	8.07	1.29	.197
Condition contrast 2 * selective attention	11.11	8.03	1.38	.166
Condition contrast 3 * selective attention	-7.63	8.09	-0.94	.346

Note. Condition contrast 1: experimental vs. control. Condition contrast 2: robot-assisted vs. tablet-only. Condition contrast 3: with vs. without iconic gestures.

6

A toy or a friend? Children's anthropomorphistic beliefs about robots and the relation with second language word learning

Rianne van den Berghe*, Mirjam de Haas*, Ora Oudgenoeg-Paz, Emiel Krahmer, Josje Verhagen, Paul Vogt, Bram Willemsen, Jan de Wit, & Paul Leseman. (in preparation).
* These authors had equal contributions.

Author contributions: R.B., M.H., O.O., E.K., J.V., P.V., B.W., J.W., and P.L. conceptualized the research; R.B., M.H., O.O., B.W., and J.W. collected the data, R.B. and M.H. analyzed the data and wrote the paper; O.O., E.K., J.V., P.V., and P.L. supervised and critically reviewed the paper.

Abstract

This current study investigated the degree to which children anthropomorphize a robot tutor and whether this anthropomorphism relates to their learning in a second language (L2) tutoring intervention. To this end, a robot perception questionnaire was administered prior to and following seven L2 vocabulary tutoring sessions with a humanoid robot. Children tended on average to anthropomorphize the robot, although there were large differences between the children. As a group, children did not significantly differ in their degree of anthropomorphism following the L2 tutoring sessions with the robot. However, children’s trajectories did: 20% of the children increased in anthropomorphism, 43% were constant in anthropomorphism, and 37% decreased in anthropomorphism. Further analyses showed that there was a weak but significant positive correlation between changes in anthropomorphism and scores on a delayed L2 vocabulary comprehension post-test. We do not know the causal direction of this relation, but our results underscore the need to consider children’s anthropomorphism when designing robot-assisted tutoring sessions.

Keywords: anthropomorphism, child-robot interaction, educational robots, second-language learning

Anthropomorphism

When interacting with a social robot, people have a tendency to attribute human form, characteristics, and/or behaviors to the robot. This phenomenon is called anthropomorphism (Bartneck, Kulić, Croft, & Zoghbi, 2009). People do not only anthropomorphize robots, but also many other non-human entities, such as animals, machines, and even natural phenomena (Caporeal, 1986), and this helps them to gain control over their environment (Duffy, 2003; Waytz et al., 2010). Anthropomorphism can be a useful mechanism in human-robot interaction (Duffy, 2003; Fink, 2012), because people evaluate robots more positively, collaborate better with them, and empathize more with robots that are more human-like or display more human-like behavior (Breazeal, Kidd, Thomaz, Hoffman, & Berlin, 2005; Eyssel, Kuchenbrandt, Hegel, & de Ruiter, 2012; Hegel, Krach, Kircher, Wrede, & Sagerer, 2008; Moon et al., 2014; Riek, Rabinowitch, Chakrabarti, & Robinson, 2009).

People do not all anthropomorphize robots to the same degree; there are differences between individuals in the tendency to attribute human qualities to nonhuman entities. One of the reasons for these individual differences is that people use their own experiences in rationalizing the actions of an object and in reasoning about its mental states (Epley, Waytz, & Cacioppo, 2007), and may thus ascribe different mental states to objects depending on their own experiences. Furthermore, people may differ in motivational needs to anthropomorphize objects, for example as a consequence of loneliness or a need for control (Epley, Waytz, Akalis, & Cacioppo, 2008). People who are dispositionally lonely are found to be more likely to anthropomorphize their pets than people who are not, and people who are in need of control are more likely to anthropomorphize unpredictable animals than people who are less in need of control (Epley et al., 2008). Thus, in human-robot interaction, the degree to which people anthropomorphize robots likely does not only depend on the type of robot used and the behavior the robot displays, but also on the specific characteristics and experiences of the person interacting with the robot.

While most robot research on anthropomorphism has focused on adults, it is not a tendency limited to adults. Children of all ages have been found to anthropomorphize robots (Beran, Ramirez-Serrano, Kuzyk, Fior, & Nugent, 2011; Monaco, Mich, Ceol, & Potrich, 2018). Both younger and older children have been found to attribute mental states to robots, even when noticing and discussing machine-like qualities such as the presence of sensors or an adult controlling the robot (Beran et al., 2011). Younger children are found to be more likely than older children to anthropomorphize robots (Beran et al., 2011; Kahn et al., 2012). They are in particular more likely to assign cognitive and affective beliefs to robots, such as the ability to remember people and understand people’s feelings (Beran et al., 2011), but they attribute fewer biological properties or livingness to robots than older children (Jipson & Gelman, 2007).

Change in Anthropomorphism

There are indications that children's perceptions or expectations of robots can change over time. Bernstein and Crowley (2008) asked children to judge different entities (including two robots) on livingness and intelligence. Children who had less knowledge about robots judged the robot more often as living than children that already had experience with robots. The latter group were more likely to distinguish robots from other entities that they already knew (e.g., things that are living) and judge robots as intelligent; however, not in a human-like manner, but in a unique robot intelligent manner. Westlund and colleagues (2016) framed a robot as either a social agent or a machine by using either inclusive language and second-person pronouns or third-person pronouns and the word ‘robot’. They assessed children’s anthropomorphism through a questionnaire both before and after playing a sorting game with the robot.

They did not find an effect of framing or having interacted with the robot on children's anthropomorphism. It is not clear from this single session study whether children's anthropomorphism is indeed not affected by interacting with robots, or whether one interaction session was not enough to change their degree of anthropomorphism.

Two studies have investigated where children focus on when interacting with a robot in order to inform the design of robots (Obaid, Barendregt, Alves-Oliveira, Paiva, & Field, 2015; Sciutti, Rea, & Sandini, 2014). These studies found that the shape of the robot was the primary focus of (young) children before they interacted with the robot, for example, the robot should have a head and arms. However, after interacting with some robots, the shape of the robot became less interesting for the children and the robot's sensory and motor properties became more important, that is, the robot's ability to feel and move. These studies did not specifically investigate how much children anthropomorphized the robot. They only looked at the role the shape, sensory, and motor properties of the robot played and how interaction with the robot changed children's expectations. Yet, the research suggests that sensory and motor properties, which can be linked to anthropomorphism, may become more important over time with increasing experience with robots.

Anthropomorphism and Learning

As discussed earlier, anthropomorphizing robots seems advantageous for human-robot interactions (Duffy, 2003; Fink, 2012), but it is not clear if and how anthropomorphism can affect robot-assisted learning. Yet, the degree to which learners anthropomorphize robots may play an important role in learning situations too, as learning is first and foremost a social process (Vygotsky, 1978). The robot's potential for social interactions to establish common ground is one of the advantages social robots in theory have over other forms of technology such as tablets. Physical robots indeed have generally been found to be more enjoyable and a preferred social partner compared to their virtual counterparts (Kidd, 2003; Pereira, Martinho, Leite, & Paiva, 2008). In theory, robots are more natural conversational partners, and they may use human-like behaviors such as gestures to support learning (de Wit et al., 2018; Macedonia et al., 2011; Tellier, 2008). This suggests that robot-assisted learning interactions benefit from similar social behaviors as human learning interactions, in particular gestures (de Nooijer, van Gog, Paas, & Zwaan, 2013; Kelly, McDevit, & Esch, 2009). Findings on the effect of this embodied presence of robots on learning as compared to virtual robots, however, are mixed (Kennedy, Baxter, & Belpaeme, 2015a; Leyzberg, Spaulding, & Scassellati, 2014).

While robots have clear advantages in theory for supporting learning, it is not clear whether the degree to which learners anthropomorphize the robot affects how much they learn from robot-assisted learning. Children who anthropomorphize the robot more might interact with the robot in a more similar way as they would interact with peers. Literature on peer learning shows the potential benefits of peers on learning (O'Donnell, 1999; King, Staffieri, & Adelgais, 1998; Topping, Hill, McKaig, Rogers, Rushi, & Young, 1997; Yarrow & Topping, 2001), and robots may have similar benefits for learning when performing a role as a peer. However, it is possible that a robot's benefits depend on the degree to which the learners anthropomorphize it. This begs the question whether anthropomorphism and learning are related to each other, which is the central research question of the current study.

Research that comes closest to answer this question is that of Chandra et al. (2018), who investigated whether children's perception of a robot in terms of intelligence, likeability, and friendliness affected their learning in a learning-by-teaching paradigm. Twenty-five seven-to-nine year old children taught a NAO robot to write over the course of four sessions as a way to improve their own writing. There were two conditions: (1) the robot improved its handwriting for half of the children, and (2) the robot did not improve its writing for the other half of the children. Children in the first condition were able to perceive the robot's improvement by the last session, but this as such did not change how they perceived the robot's intelligence, likeability, and friendliness. However, children's own improvement in writing was positively correlated with the likeability of the robot. In the condition in which the robot did not improve, children's perceptions of the robot's intelligence, likeability and friendliness did not change either, but in this condition children's own learning was correlated with the perceived friendliness of the robot. These findings need to be interpreted with caution because of the small sample size, but they suggest that children's perception of the robot may indeed be related to their learning.

Our study expands previous work in two ways. First, our study includes a larger sample in a multiple sessions second-language (L2) learning experiment. Second, we assess the degree to which children anthropomorphize the robot both before and after having interacted intensively with it, allowing to observe changes in anthropomorphism and to examine the relations between children's anthropomorphism and changes therein with language-learning gains.

This Study

The current study was conducted within the L2TOR project, which is a research and development project on the use of social robots in young children's L2 learning (Belpaeme et al., 2015; Belpaeme et al., 2018). The current study was part of a large-scale randomized controlled trial within the L2TOR project to evaluate the effectiveness of a multiple sessions L2 learning intervention for young children using a social robot. The study included four conditions: (1) an L2 vocabulary training with a tablet and a robot that performed iconic and deictic gestures to support word learning (gestures that visualize target words and pointing gestures), (2) an L2 vocabulary training with a tablet and a robot without iconic gestures (only pointing gestures), (3) an L2 vocabulary training in with a tablet only (no robot involved), and (4) control condition in which children only played dancing games with the robot. In the current study, we only included the experimental robot conditions to investigate children's perception of the robot and the way in which their perception of the robot relates to their learning. We address the following research questions and hypotheses:

1. To which degree do children anthropomorphize the robot? We expected children to differ in the degree they anthropomorphize the robot, in line with research on individual differences in anthropomorphism (Epley et al., 2008).
2. Do children's perceptions change through multiple L2 tutoring sessions with the robot? Although the evidence is mixed (e.g., Bernstein & Crowley 2008; Kory Westlund et al., 2016), we expected that children's perceptions would change over time in different ways, due to the multiple interactions children have with the robot. On the one hand, children may come to perceive the robot more as a friend after repeated interactions, thus perceive the robot as more human-like. On the other hand, it is also possible that children have initially high expectations of the robot's interactive qualities, which the robot, however, cannot meet. In that case, their perception would change to a perception of the robot as less human-like.
3. Are children's perceptions of the robot related to their learning of L2 words? We expected that children who anthropomorphized the robot more would perceive the robot more as a peer learner throughout the tutoring sessions than children who anthropomorphized the robot less. As learning is a social process (Vygotsky, 1978), we expected children's learning to benefit from perceiving the robot as a peer learner, in line with literature on the benefits of peers on learning (O'Donnell, 1999; King et al., 1998; Topping et al., 1997; Yarrow & Topping, 2001).

Method

Participants

This study reports on a part of the sample described in Chapter 5. Data was used from 108 monolingual Dutch children (48 girls) with an average age of 5 years and 8 months (SD = 5 months) who followed the vocabulary training in one of the two robot-assisted conditions (with or without iconic gestures). These children were recruited from the kindergarten departments of nine primary schools in the Netherlands. Within schools, children were randomly assigned to one of the conditions, while ensuring a similar gender distribution over the conditions. There were 54 children (22 girls) in the iconic-gesture condition (*M* age = 68.4 months, *SD* = 4.8 months) and 54 children (26 girls) in the no-iconic-gesture condition (*M* age = 68.5 months, *SD* = 4.7 months). Twelve additional children were initially included and pre-tested, but did not complete the experiment due to technical issues or because they did not want to participate anymore. All children's parents signed an informed consent form to allow their children to participate in this study.

L2 Tutoring Sessions

The aim of the L2 tutoring sessions was to teach each child 34 English words in the domains of mathematical and spatial language. Each child received seven tutoring sessions involving the robot and a tablet. The Softbank Robotics NAO robot was used, which was sitting in a 90 degree angle next to the child (see Figure 1).

The tablet taught the robot and the child the target words. For each word, the child and the robot had to perform different tasks on the tablet (e.g., dragging objects on the screen, repeating target words, or acting out target words) or to act out target words. During these tasks, the robot acted as a slightly more knowledgeable peer who was also taught English, but could provide feedback on the child’s actions when

Figure 1. A child playing with the robot.

needed. For example when a child was reluctant to drag an object on the tablet, the robot could perform this task for the child. In the iconic-gestures condition, the robot used an iconic gesture for every target word, while, in the other condition, it did not. All other gestures were exactly the same across conditions. The complete interaction was autonomous, except for the recognition of children's speech. The interaction was a one-on-one interaction, but the experimenter stayed in the same room to intervene when necessary. During each of the sessions children were introduced to five or six new target words. Prior to the first session, they received a pre-test to assess their knowledge of the English target words. An immediate post-test was administered within two days of the last session, and a delayed post-test two to five weeks after the immediate post-test using a receptive comprehension task to determine the learning gains (for more details on the study, see Chapter 5). In addition to the language test, children's anthropomorphism was assessed twice, before and after the lesson series.

Materials and Measurements

In this chapter, we focus on two different constructs: 1) the degree to which children anthropomorphize the robot as measured with a perception questionnaire, administered prior to the first tutoring session and after the seventh and last session; 2) children's L2 vocabulary learning as measured with a comprehension test, administered at an immediate post-test and a delayed post-test. Other measurements were taken as well, but are for reasons of brevity beyond the scope of this chapter as they assessed other variables predicting children's learning gain but not anthropomorphism (see Chapter 5; Vogt et al., 2019 for more results).

Perception questionnaire

This perception questionnaire was constructed for the purpose of the present study and administered by an experimenter in a one-by-one session with the child. The questionnaire took about ten minutes to complete. It consisted of twelve questions (for an overview, see Table 1 in the Result section) and assessed children's perception of the robot with regard to various types of properties: biological (e.g., feeling pain, having to eat, and growing), cognitive (e.g., thinking, remembering), and emotional (e.g., being happy, being sad). Each question could be answered with 'yes'/'no'/'I don't know' and was followed by an open-ended query asking children why they gave this response. The items were based on Jipson and Gelman (2007) who investigated to what extent children make a distinction between living and non-living items. Two additional questions were included in the questionnaire ("can break" and "is made by humans") but proved unreliable, as children's answers to the open-ended query did not correspond to their answers on the close-ended questions. Therefore, we removed the answers on these questions from the data. The children were awarded one point for each 'yes' answer. Thus, the maximum score was twelve, with a higher score denoting a child's tendency to consider the robot as human-like. Cronbach's alpha indicated that the internal consistency of the questionnaire was satisfactory, α = .72 at the pre-test and α = .75 at the post-test.

Comprehension test

The comprehension test was a picture-selection task. In this task, children were presented with a prerecorded target word and asked to choose which one out of three pictures or short video clips matched this word best ('Where do you see: [heavy]?'). Each target word was presented three times with different distractors in a random order. Only half of the 34 target words that were presented in the vocabulary training were included, as a test including all target words would have been too long for these young children. The same test was used at both post-tests. The internal consistency of the comprehension task was good, with Cronbach's alpha α = .84 at the first post-test and α = .87 at the second post-test.

Procedure

Prior to the experiment all children participated in a group introduction with the robot to familiarize the children with the robot, build trust, and explain the basic similarities and dissimilarities between the robot and humans (e.g., the robot speaks without moving its mouth, but looks at us while speaking in the same way as humans do; Vogt, de Haas, de Jong, Baxter, & Krahmer, 2017). These explanations were deemed necessary to make sure that children would know how to interact with the robot in the subsequent lessons. During the introduction, participants danced together with the robot, were allowed to shake the robot's hand, and played a brief gesture imitation game. The robot was not explicitly framed as either a robot or a machine, by avoiding pronouns and by being called 'Robin the robot' (i.e., a combination of a gender-neutral human name and the label 'robot'). After the introduction, the first perception questionnaire was administered, together with other measurements. In the weeks thereafter, the children received seven one-on-one tutoring lessons with the robot, after which the perception questionnaire was administered for the second time, together with the immediate comprehension post-test. Finally, the comprehension test was repeated at a delayed post-test, between two and four weeks after the lesson series ended.

Data Preparation and Analyses

Children's score on the perception questionnaire was the number of points awarded, divided by the number of questions that were administered. We used proportions rather than total scores because there were missing values on some items for some children. This was the case for one child at the pre-test (four of the twelve questions were not administered) and for five children at the post-test (for each of whom one question was not administered). Furthermore, complete pre-test data of three children and post-test data of one child were missing due to illness. Note that we do not discuss gender in the Result section. We ran our analyses with gender as an additional independent variable, but did not find any gender effects.

Results

Anthropomorphism of the Robot

First, we investigated to which degree children anthropomorphize the robot (RQ1). Table 1 displays the questions of the questionnaire and the average scores of the children, with the standard deviations revealing large differences between the children. As a group, children tended to anthropomorphize the robot as is reflected in the overall scores being higher than .50 at both the pre-test and the post-test, but the scores varied strongly between the questions. Children highly agreed that the robot 'can enjoy something', 'can be happy', and 'can think'. They disagreed more often on various biological properties, such as 'feeling it when you tickle Robin the robot' and 'feeling pain'.

Table 1. Mean Proportions (SD) on the Questionnaire during the Pre- and Post-Test

Do you think that Robin the robot…	Pre-test	Post-test
… can see things?	.81 (.42)	.80 (.41)
… can be sad?	.69 (.48)	.45 (.52)
… can remember something?	.67 (.49)	.76 (.49)
… can feel it when you tickle Robin the robot?	.56 (.59)	.44 (.60)
… can think?	.80 (.43)	.78 (.54)
… has to eat?	.29 (.48)	.21 (.45)
… understands when you say something?	.72 (.49)	.89 (.50)
… can feel pain?	.46 (.50)	.39 (.58)
… can enjoy something?	.96 (.28)	.96 (.27)
… grows?	.17 (.40)	.16 (.41)
… can be happy?	.96 (.24)	.99 (.35)
… can recognize you?	.64 (.62)	.94 (.33)
Overall scores	.60 (.19)	.58 (.20)

Change in Anthropomorphism

Second, we investigated whether children’s perception had changed after the L2 tutoring sessions with the robot (RQ2). There was a moderately strong correlation between pre-test and post-test scores on the perception questionnaire, $r(105) = .505$, $p < .001$, indicating moderate overall stability of children’s perceptions. However, as Figure 2 shows, there was also large variability among the children in whether and how their perceptions changed between the pre- and post-test. Most children were consistent in the degree to which they anthropomorphized the robot (45 children), but a relatively large number of children anthropomorphized the robot less after having interacted with it in the tutoring sessions (35 children). An increase in anthropomorphism also occurred, but was least common (24 children).

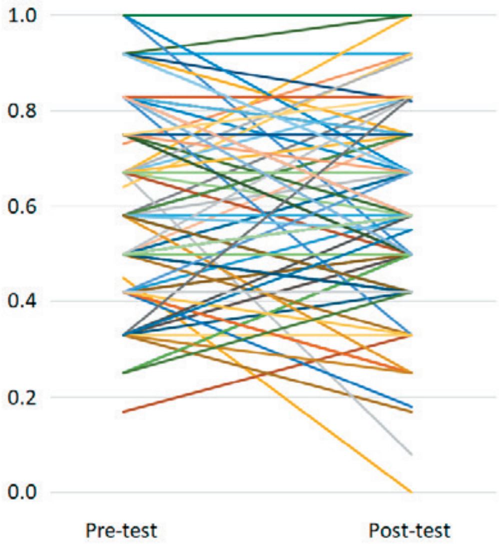


Figure 2. Change in children's proportion scores on the perception questionnaire.

We compared children's answers on the perception questionnaire at the post-test to those at the pre-test. A paired samples t-test did not show significant differences between children's overall pre- and post-test scores on the perception questionnaire, $t(103) = 1.53$, $p = .130$, $d = .10$. However, Table 1 shows that children changed their opinion drastically on a number of questions. Fewer children believed at the post-test that the robot could feel it when being tickled, that it could feel pain, or that it could be sad. More children believed during the post-test that the robot could understand what they said, and that the robot could recognize them.

We explored whether children perceived the robot differently in the iconic-gesture condition compared to the condition without iconic gestures, using an ANOVA. There were no differences between the two conditions in the degree to which children anthropomorphized the robot, $F(1,102) = .00$, $p = .957$, $\eta_p^2 = .00$, and condition did not interact with time, $F(1,102) = .64$, $p = .424$, $\eta_p^2 = .01$. Thus, the use of iconic gestures apparently was not associated with a higher degree of children's anthropomorphizing of the robot.

Anthropomorphism and Learning

Lastly, we investigated relations between children's perception of the robot and their learning gains over the tutoring sessions (RQ3), using Pearson's correlations (see Table 2). Pre-test anthropomorphism was weakly related to the comprehension scores on the immediate post-test, *r*(104) = -.21, *p* = .034. The relation was negative, suggesting that children who anthropomorphized the robot more prior to the lesson series learned less than children who anthropomorphized the robot less. Post-test anthropomorphism was not related to comprehension scores on either post-test, both *p*s > .100.

Children's change in anthropomorphism was weakly but significantly related to the comprehension scores on the delayed post-test, *r*(104) = .21, *p* = .031. Thus, an increase in the degree to which children anthropomorphized the robot was associated with higher performance on the delayed comprehension test.

Table 2. Correlations between the Anthropomorphism Scores and the L2 Comprehension Scores

	Comprehension	
Anthropomorphism	Immediate post-test	Delayed post-test
Pre-test	-.208*	-.137
Post-test	-.152	.094
Change	.036	.212*

Note. * *p* < .05.

Discussion

In the present study, we (1) investigated the degree to which five-year-old children perceived a social robot as human-like, (2) whether this perception changed after intensive experience with the robot acting as a peer tutor in an L2 word learning intervention, and (3) whether this perception and the change therein were related to children's learning gains.

Anthropomorphism of the Robot

We investigated the way children perceived the robot after a group-wise introduction session and prior to the tutoring sessions. Overall, children slightly more often agreed than disagreed with statements attributing human-like properties to the robot, but there were large differences between children in their perception of the robot, in line with research on individual differences in the tendency to anthropomorphize objects (Epley et al., 2007; Epley et al., 2008). Moreover, children agreed more often with statements that attributed cognitive and, to some extent, also positive feeling states to the robot than biological properties and negative emotional states, in line with previous work that also found that young children are likely to ascribe cognitive mental states to robots (Beran et al., 2011).

As this was not the scope of the current study, we did not present and analyze children's answers to the open-ended questions, which asked them to motivate why they perceived the robot as more or less human-like. However, we noticed that there were large differences between the children, similar to their perception scores, in the way they explained why they perceived the robot in the way they did. For example, some children thought that the robot would be sad if children did not want to play with it, while other children thought the robot would be sad if it was in pain. Some children thought that the robot could not be sad because it had no feelings, while other children thought the robot could not be sad because it could not handle water and, thus, could not cry.

Change in Anthropomorphism

We investigated whether children's perception of the robot had changed after the L2 tutoring sessions. There were no significant differences in children's overall perception of the robot and also at the post-test children on average slightly more agreed than disagreed with attributing human-like properties to the robot. However, with regard to specific properties some major changes were observed. Fewer children answered 'yes' to questions attributing biological properties and negative emotions to the robot at the post-test as compared to the pre-test. This concerned, for example, questions asking whether the robot 'could feel it when being tickled' or 'could feel pain'. This is in line with the study of Sciutti et al. (2014) in which was found that the robot's sensory and motor properties became more salient to children after they had interacted with a robot. At the post-test, more children answered 'yes' to questions addressing whether the robot can remember something, understand them when they say something, and is able to recognize them.

These changes together indicate an interesting shift in the way in which the robot is seen by children after intensive experience, namely as a basically mechanical being but with positive mental states, whereas initially children showed more confusion regarding the biological aspects and were less strongly convinced of the cognitive capabilities of the robot. We believe that this shift is due to the way in which the lessons were designed. At the start of each lesson, the robot greeted the children personally while mentioning their names, referred to the previous lessons and tracked the children's faces to suggest that they looked at the children. It is likely that children were less inclined to believe that the robot could recognize them at the pre-test, simply because they had not yet played intensively with the robot in a one-on-one setting yet at that time. Regarding negative emotional states, fewer children believed at the post-test that the robot 'could be sad', which can also be explained by the design of the lessons. Even though the robot expressed happiness (by changing the colors of its eyes) and, therefore, also when it was *not* specifically happy (the colors of the eyes did not change), it did never express negative emotions, like sadness or anger.

Most children either anthropomorphized the robot either to the same degree or to a lesser degree during the post-test as compared to the pre-test. Fewer children increased their anthropomorphism of the robot. It is possible that decreases in anthropomorphism were due to children having high expectations of the robot's interactive (human-like) qualities, which the robot could not meet (Dautenhahn & Werry, 2004). The robot was largely autonomous during the tutoring sessions and did not engage in personalized conversations with the children. The robot kept to the preprogrammed script and did not answer children's questions. For children with high expectations regarding the human-likeness of the robot, this could have led them to decrease their attribution of human-like properties to the robot. Conversely, children who had a less human-like perception of the robot prior to the tutoring sessions may have had low expectations of the robot's interactive (human-like) qualities. Since the robot displayed at least some human-like behaviors, such as mentioning the child by name (suggesting that it recognized the child) or indicating that it liked the sessions, this could have increased children's beliefs about the robot as human-like over repeated interactions. Thus, the observed changes in perception of the robot may have been more dependent on children's prior expectations rather than the robot's design and behaviors.

A final possibility is that the observed change in anthropomorphism merely reflects the phenomenon of regression to the mean, with initially higher scores decreasing and initially lower scores increasing at post-test due to random measurement error. While we cannot fully rule out this explanation, it should be noted that more children decreased than increased in anthropomorphism, whereas the analysis at the item level revealed a complex but interpretable pattern of changes that pointed to a shift in how children perceived the robot within a similar overall anthropomorphism score at the pre- and post-test.

Anthropomorphism and Learning

Finally, we investigated whether children's perception of the robot was related to their learning gains. We found two weak but significant correlations. Children's anthropomorphism of the robot at pre-test was negatively related to their comprehension scores at the immediate post-test, though not at the delayed post-test. In contrast, a change in perception towards more anthropomorphism was positively related to learning gains at the delayed post-test, though not at the immediate post-test.

Against our expectations, only a change towards more anthropomorphism was positively related to word learning, and not children's pre- and post-test anthropomorphism. Possibly, this points again to the role of children's expectations about the robot as a human-like being. If children had high expectations which the robot could not meet, they may have become disappointed while working with the robot over several tutoring sessions. As a likely consequence, they may have become less engaged, which is not beneficial for learning. In contrast, for children with low expectations, the robot exceeding these expectations may have had a positive effect on their engagement during the tutoring sessions and, through this, on their learning.

There are two important caveats. First, the correlations though statistically significant were rather weak. Moreover, we did not include child characteristics such as age and cognitive ability that could possibly underlie the observed correlations. It is possible that the correlations are spurious and caused by a shared third factor.

Second, the present design did not allow to test the causal direction of the observed correlations. Thus, it is not clear whether children learn more from the robot because they come to perceive it as a human, or that they come to perceive the robot as more human-like because they have successful language-learning interactions with it.

Limitations, Strengths, and Future Research

The current study has several limitations. We did not use a standardized questionnaire for anthropomorphism because of our young target group. Standardized tests such as the Godspeed questionnaire (Bartneck et al., 2009) often use Likert scales which are too difficult for young children. However, we based our questionnaire on previous work (Jipson & Gelman, 2007) and the questionnaire proved to be reliable, showing also moderate stability between pre-test and post-test. Furthermore, we do not know how the group-wise introduction of the robot before the pre-test affected children's perception of the robot. To ensure that children could establish a common ground with the robot and to avoid anxiety, the introduction contained several statements about the properties of the robot, for example, being a peer, speaking as a human and looking as a human, that may have biased children's perception towards anthropomorphism at the pre-test. However, administering the perception questionnaire prior to the introduction, would have had other disadvantages. For instance, it would not have been clear whether children's perceptions were based on actual interactions with similar robots, with different robots, or were based on cartoons, movies or television programs, or just on imagination. The large variation in scores indicates that children still formed their own opinions about the robot, but we do not know whether these opinions were biased towards anthropomorphizing. Note that despite this possible bias, the *changes* in perception we observed, in particular at the item level, can be considered genuine and likely to relate to the intensive experience children had with the robot during the lessons. Finally, we could only conduct correlational analyses to examine how children's perception and learning were related. Moreover, we could not rule out that other child-related factors underlie the relations that were observed between children's anthropomorphism and learning gains. Future research with experimental designs is needed to test whether framing the robot as a machine or as similar to a human affects children's learning differentially. A high level of anthropomorphism in itself may not be required for successful tutoring sessions, as no positive main effects of anthropomorphism were found in our study. Managing children's expectations of robots, on the other hand, may be important, as changes in anthropomorphism could relate to learning gains.

Our study has also several strengths. It is one of the first studies to investigate changes in children's anthropomorphism after multiple exposures to a robot. It is also the first to explore how anthropomorphism and changes therein as a consequence of intensive interaction with the robot relate to children's learning. Moreover, the different robot properties presented in the questionnaire allowed for a more thorough and differentiated understanding of the ways in which children perceive robots.

Conclusion

The study presented in this paper explored how children anthropomorphize a humanoid robot, whether their perception had changed after seven tutoring sessions, and whether the change in perception correlated with children's learning gain during these sessions. We found that children generally anthropomorphize the robot, although there were large differences between children in the degree to which they did. Our results showed that children's overall tendency to anthropomorphize had not significantly changed after the tutoring sessions, but the analysis at the item level revealed a complex pattern of changes indicating a shift within this overall tendency towards seeing the robot as more mechanical while at the same time attributing more cognitive capabilities to the robot. As an exploration, we found a weak but significant correlation between children’s increased anthropomorphism and their word learning gains. Children who came to perceive the robot more as a human learned more from the tutoring sessions. Although the causal direction of this relation is not yet clear, the results underscore the importance of taking children's anthropomorphism into consideration when designing robot-assisted tutoring sessions.

7

General discussion

The increasing linguistic diversity of society challenges education. On the one hand, there is a growing need to provide second-language (L2) education to young children at an increasingly young age. On the other hand, home languages of children who grow up with a language different from the (pre)school language need to be supported. Whereas there is strong evidence for the effectiveness of building upon children's first language (L1) when learning the L2, the language of the school, and allowing children to use their L1 in school, there are many practical obstacles to supporting children's L1s that become even more pressing if more languages are represented in a classroom. Technology, and specifically social robots, may aid in tackling some of these challenges. With their humanoid appearance and presence in the real world, robots can stimulate natural and embodied interactions better than other forms of technology, in line with current understanding of language learning as embodied and embedded to ground meaning (Barsalou, 2008; Hockema & Smith, 2009; Iverson, 2010; Wellsby & Pexman, 2014). Moreover, it is possible, at least in theory, to program robots as speakers of virtually any combination of languages.

Social robots increasingly find their way into (L2) education, but little is known yet about their effectivity (see Chapter 2; Kanero et al., 2018, for reviews on social robots for language learning). Specifically, much is still unclear regarding how to design human-robot interactions for teaching purposes and how particular design choices affect the effectiveness of social robots in (language) learning. The current dissertation was carried out within the L2TOR project, which had the goal to tackle some of these issues raised above by developing a robot that could aid in children's L2 learning. To this end, we designed a series of seven lessons in the domains of mathematical and spatial language to teach Dutch children new words in English in these domains. The lesson series resembled word learning programs that were proven to be effective, yet involving a human tutor. The lessons focused on domain-specific academic words in the area of mathematical and spatial language, also referred to as tier-two words (Beck & McKeown, 1985), 34 in total. These words were embedded in age-appropriate narratives and taught by a robot tutor. A major task within the L2TOR was the evaluation of the lesson series and the robot tutor. To assess the effectiveness of the robot tutor, a randomized controlled trial was conducted with five-year-old children.

Already at the start of the project, it became clear that many technological issues had to be resolved or circumvented before a realistic L2-word learning intervention could be implemented. Speech recognition, an essential first step to create an interactive robot, turned out to be far from perfect for use with young children (Kennedy et al., 2017). Furthermore, the robot was neither able to recognize objects nor to manipulate them, limiting the possibilities to work with physical objects or to act out particular operations, such as addition and subtraction. We had to work around these limitations by including a tablet as a mediating device in the setup such that children could perform actions on 3D representations of objects presented on the tablet's screen, and the robot could use the tablet data as input and respond to children's actions.

Other technical limitations concerned the lack of interactivity and adaptivity of the robot, and the possibilities to provide tailored feedback to children: All actions and utterances of the robot had to be programmed beforehand. Although children at this young age are still highly idiosyncratic in their learning routes and still rather unpredictable in their behavior, the tutoring situation had to be standardized and protocolled. This resulted in a vocabulary training that did not reach the same naturalness, flexibility, and adaptivity as would have been obtained when a sensitive human tutor was involved. Working around all these limitations, we succeeded to create a series of lessons that could be applied in a school situation with a sufficient degree of ecological validity, providing the opportunity to test the potential of a social robot as an L2 tutor.

The results of the randomized controlled trial are reported in this dissertation (see Chapters 5 and 6) in addition to the findings of the studies that were conducted prior to this main study to inform the design of an effective RALL interaction (see Chapters 3 and 4). The aims of this dissertation were (1) to synthesize earlier findings on the use and effectiveness of social robots for language learning, and identify in particular current issues and avenues for future research; (2) to investigate the use of tablets to inform the design of RALL interactions; (3) to investigate the added benefits of robots for L2 word learning; and (4) to investigate whether a robot's benefits depend on individual differences in children's language learning and attention skills, and their perception of the robot. A review and four experimental studies were conducted, to address these aims. Below, we discuss the main findings of the studies, and reflect on the future of robots in (language) education.

Previous Work on RALL

In Chapter 2, a narrative review was presented on the current body of literature on social robots as used for language learning. Thirty-three studies on RALL, targeting different languages, age groups, and aspects of language, and using different robots and methodologies, were discussed. Mixed results were found with respect to L1 and L2 learning outcomes. For word learning in particular, current research has not convincingly shown that robots are effective. Their effects appear to depend on their role (e.g., teaching assistant or peer learner), their target age group (e.g., preschool children, school-aged children, or adults), and the number of sessions (one or more). The few studies examining reading skills, grammar learning, and sign language showed quite positive results, while the evidence with respect to speaking skills was more mixed. As discussed, differences in demands on the robot's interactional qualities (e.g., being able to have contingent conversations) likely differ between lessons on speaking skills and lessons on reading or grammar, and may explain these mixed findings. Mediating devices displaying educational content can easily be used in lessons on reading and grammar, while the robot needs more conversational skills to teach speaking skills. More positive results were found regarding participants' learning-related emotions, as both children and adults were often found to enjoy working with the robot. However, these positive emotions could be short-lived and disappear upon more frequent interaction with the robot, a phenomenon that was found in a number of studies. The conflicting results with respect to the social behavior of the robot highlight the difficulties of designing effective RALL interactions. Some studies found positive effects of personalized and/or social behavior on learning gains and enjoyment, while other studies found social behavior to negatively affect these outcomes. Future research will have to investigate how to design effective RALL interactions, and, as we will argue below, maybe they should shift the focus from attempts to mimic human tutors to capitalizing on the strengths of robots to complement the human tutor. One such design issue, namely the use of tablets, was addressed in the study presented in Chapter 3, discussed below.

Designing RALL Interactions

The underlying assumption of most RALL studies is that these child-robot interactions should resemble human-human interactions as much as possible. As discussed above, when the robot's skills fail to mimic human interactional skills (e.g., speech recognition) the effectivity of the robot may decrease. However, it is not yet known if children interact with social robots in the same way as they interact with humans. In this dissertation, we examined an aspect known to be important for language learning in human-human interaction and the way in which it could be implemented robot-child interaction, namely interaction with physical objects. Chapter 3 presented a study on the use of tablets rather than physical objects. As discussed, a tablet had to be used in the L2TOR system because the robot was not able to recognize or manipulate objects. A concern was whether this would hamper children's learning as compared to the use of physical objects. Therefore, the use of physical versus virtual object in an L2 vocabulary training was compared. Following an embodied cognition perspective (Barsalou, 2008; Hockema & Smith, 2009; Iverson, 2010; Wellsby & Pexman, 2014), we predicted that physical objects would, in contrast to virtual objects on a tablet screen, activate children's underlying embodied concepts through sensorimotor interactions, and thus benefit their learning by grounding the new words in these underlying concepts. However, children who manipulated physical objects did not outperform children who manipulated virtual objects on a tablet screen on any of the word-learning tasks at both an immediate and a delayed post-test. We do not claim that tablets are just as effective learning tools as physical objects for all learners and in all domains, but manipulating 3D objects on a tablet did not appear to affect preschool children's L2 word learning differently than physical objects. Perhaps any manipulations (including swiping on a tablet screen) can benefit learning, or manipulating physical objects may be less important when learning L2 words than when learning L1 words for which the learner has no concept yet. The use of physical objects may particularly benefit learning if the learner still has to acquire the underlying, embodied concept. In summary, using tablets as mediating devices instead of physical objects does not necessarily hamper children's learning in child-robot interactions.

The Added Value of Social Robots and Individual Differences between Children

The studies presented in Chapters 4 to 6 investigated the added value of social robots for language learning and examined whether the effects of using a social robot differed between children. In the study reported in Chapter 4, children taking an L2 vocabulary training with the robot who took the role of a peer were compared to children who took the training either without a peer or with a child peer. In contrast to our expectations, children in the three conditions did not perform differently on an immediate post-test, and children being taught English words without a peer even outperformed children in the child-peer and robot-peer conditions at the delayed post-test. We also did not find that children in the three conditions differed in the degree in which they enjoyed the training, and differences in enjoyment did not impact children's learning. Perhaps child-robot interactions need to be more interactive to ensure that robots can actively support children's learning in order to find benefits of robots on word learning.

In the large-scale randomized controlled trial reported in Chapter 5, in which children participated in a comprehensive multiple sessions L2 vocabulary training on a tablet either assisted by a robot or not, children were again found not to benefit additionally from being assisted by the robot. Children received an L2 vocabulary training either with a tablet-only, with a robot that used deictic gestures, or with a robot that used deictic and iconic gestures. They were compared to a control group of children who danced with the robot and were not taught any L2 words. Children learned from the vocabulary training. Children in the experimental conditions outperformed children in the control condition on vocabulary measures. However, there were no differences in learning gains between the tablet-only, no-iconic-gestures and iconic-gestures robot conditions.

In addition, the analyses of individual child characteristics showed moderating effects of children's pre-intervention language learning and attention skills on L2 vocabulary learning. Larger L1 vocabulary and stronger phonological memory benefited learning L2 words in the robot without iconic gestures condition, while a higher level of selective attention particularly benefited learning in the robot with iconic gestures condition, suggesting that good attention skills are required to benefit from the robot's gestures. Children with larger L1 vocabularies and/or better phonological memory benefited from the robot's presence, while they apparently did not need its iconic gestures to learn the target words.

A likely explanation for the general lack of learning benefits of the robot is that children were highly focused on the tablet in these lessons. The tablet was required given the current state of technology, but led to an interaction that revolved more around the tablet that displayed the educational content than around the robot. The robot's physical presence and its possibilities for play with children are the most important advantages of robots over other forms of technology such as tablets. These advantages currently mostly exist in theory but cannot be implemented in practice yet. Moreover, the moderator effects of the language and attention skills show that RALL may not be suited for all children equally and that individual differences in children's skills should be taken into account when designing RALL interactions.

Children's perceptions of the robot were explored in a final study presented in Chapter 6. A questionnaire measuring the degree to which children anthropomorphized the robot (i.e., attributed human-like characteristics to the robot) was administered prior to and after the seven L2 vocabulary tutoring sessions. Children were generally found to anthropomorphize the robot, but there were large individual differences. Although the overall degree of anthropomorphism did not change between pre- and post-test, shifts at the item level indicated that children saw the robot more clearly as a mechanical being with cognitive mental states instead of as biological being that could have negative emotional states. In addition, a weak but significant positive correlation was found between a change towards more anthropomorphism (indicating that children saw the robot as more human-like after the lessons series) and the scores on a delayed L2 vocabulary comprehension post-test. We did, however, not find a main effect of anthropomorphism, indicating that anthropomorphism in itself may not be required for successful tutoring sessions, at least not in relatively simple, highly standardized word learning lessons. Managing children's expectations of robots, on the other hand, may be crucial, as a positive change in anthropomorphism was related to larger L2 learning gains. Children who learned more words thus were also the children who attributed more human-like characteristics to the robot after having interacted with it. Other children, however, may have had high expectations of the robot which the robot could not meet, and those children learned fewer words. Although we cannot tell the causal direction of this relation, it is clear that (managing) children’s perception of the robot is worth investigating further.

The Future of Robots in (Language) Education[10]

Further technological developments are clearly necessary before robots can be used to support language learning in education in a way that does justice to their potential advantages. For example, speech recognition and object recognition are needed to develop interactions in which robots can, at least to some extent, understand children and play with physical objects. Moreover, recognition of emotions and non-verbal behavior is needed to monitor children’s emotions and engagement during the training, and to adapt the training accordingly (e.g., re-engaging children when they become bored or distracted). If these technological requirements are met, lessons can be developed that make better use of the potential advantages of robots. To work around the technical limitations, we designed a system in which the robot was very

[10] A previous version of this part was included in the L2TOR EU report *D7.6 Integrated report and recommendations* (van den Berghe, Kramer, et al., 2018).

static. It followed predefined scripts, in which researchers and developers had invested many hours to develop. The robot was not flexible. It could not divert from its script to adapt to the situation. For example, the robot did not change the way in which it explained the meaning of a certain word depending on the responses and actions of the child. If a child did not give the correct response, predefined feedback was provided. Moreover, the robot could not tell when children were distracted or needed a different way of tutoring. It could not sense children's input, and even if it could, it would not 'know' how to deal with it. Within the L2TOR project, researchers have focused on trying to make the robot adaptive, that is, to try to adjust the difficulty of the lesson to the learner's knowledge using sophisticated Bayesian models (Schodde et al., 2017; de Wit et al., 2018). However, even these adaptive robots appeared to be limited in their possibilities. Robots cannot, given the current state of technology, adapt as much as is needed to create contingent RALL interactions. They cannot yet simultaneously monitor the learner's knowledge, mental state, emotions, and movements, and adapt their own behavior accordingly.

Having stated all this, we do believe that robots have potential and we expect robots to become part of the educational landscape in years to come, although perhaps in a different way. Below, we present some suggestions for how robots can, in the future, be implemented in educational contexts. Perhaps robots need to be much more intelligent to truly harbor their potential. There have been major developments in the field of artificial intelligence in recent years, and robots until now rarely incorporate the more advanced artificial-intelligence systems. This could, and perhaps should, be changed.

In their seminal paper, Smith and Gasser (2005) discuss six lessons learned from the development of human infants that should, in their view, guide the development of embodied intelligent agents (usually taken to imply AI systems). Perhaps robots need to go beyond being a physical body with simple computers in it to entities with artificial-intelligence systems that have a sort of embodied intelligence. The six lessons Smith and Gasser draw from babies are the following, in short: be multimodal (i.e., have concepts that are intrinsically grounded in and defined by coordinated multiple sensory and action schemes), be incremental (i.e., learn), be physical and explore (i.e., learn about the real world in real time), be social (i.e., be empathetic and learn about social rules), and learn a language (which should not only be about word-word relations, but also about word-world relations; cf. Pulvermüller, 2013). For the remainder of this section, we will assume that it is possible to develop an embodied intelligent agent according to these recommendations, at least to some extent. Below, we discuss each of these six lessons and describe how a robot as a language tutor would benefit from being an embodied intelligent agent. It seems clear that not all recommendations can be equally easily followed-up due to hardware constraints and other technological issues.

Lesson 1: Be multimodal

The first lesson concerns multimodality. Children learn through the various ways in which they come into contact with the environment, such as vision, audition, touch, and smell. They learn that their sensory systems are interrelated and the primary concepts they develop about the world consist in coordinated multimodal sensorimotor schemes. For example, the perception of an object changes if it is grabbed and moved, while at the same time the time-locked coordination of the varying perceptions with the motor movements underlie the integrated perception of invariant structure, which is the basis of multimodal object knowledge in the human infant. In our current robot, the robot uses few sensory systems and the different systems are not truly interrelated. Moreover the knowledge of the robot is essentially amodal and abstract (e.g., visual input is translated into a general information format, loosing much of its modality-specific richness). The robot in our experiments received input from only the tablet and its own cameras (which the robot could only use for face tracking but not for other types of vision such as object recognition). A robot that would have multiple sensory systems which it could integrate and relate to movement information in real time, would be a very different robot tutor. This robot would be able to perceive objects as invariant structures despite the ever changing perceptions when manipulating objects or when moving around, it would create concepts which are grounded in real-life experiences with objects, and it would be able to perceive and act-upon objects as they are presented in a real-time situation. As a result, the robot's gestures would also be grounded in its experiences with objects. The current way of developing robot gestures is a time-consuming procedure of modelling gestures after how the programmer thinks a gesture should look like. However, gestures may be much more subtle and grounded in one's own experiences. A robot that would have held a heavy object in reality, could subsequently gesture "holding" or "heavy" according to its own experiences with holding heavy objects. This would enable the robot to also produce a more varied repertoire of gestures for the same word that contain certain invariant aspects that are essential to conveying the meaning of this word. Such varied gestures, similar to the way humans produce gestures (Kita & Özyürek, 2007), might facilitate word learning more than the type of gestures used in this study.

Lesson 2: Be incremental

The second lesson concerns incremental learning. Currently, the robot is quite static and not learning. First steps are made towards adaptive robots, as discussed above, but the extent to which robots are really incremental is limited. An incremental robot tutor could learn from its interactions with the child and adapt the difficulty of its lessons based on the child’s current needs in the concrete instruction situation. A beginner learner may need a “simpler” robot, which does not display too many complex social behaviors compared to a more advanced learner. The learning robot can incrementally add behaviors as the learner progresses. This would also likely counter the novelty effect – the often observed decline in motivation and interest of the child in interacting with the robot. Previous research has found that a robot that adds new behaviors over time results in child-robot interactions of higher and more enduring quality than predictable robots without new behaviors (Tanaka, Cicourel, & Movellan, 2007). An incremental robot is less prone to children losing interest in the robot after having played with it for a longer period of time.

Lesson 3: Be physical

Lesson three is to be physical. Infants learn through interacting with physical objects and by linking objects, locations, and space. They can even learn words for objects that are not visible anymore while being labelled, simply by linking the label to the location in which the object was visible initially. Within the current state of technology, the robot cannot really interact with the environment. It can move itself through space, but it does not perceive the environment while moving and has no spatial representation of its actions, nor of the perspective of the interaction partner. The robot in our study could manipulate the tablet, not by physically manipulating it externally, but through internal codes that moved objects on the tablet while the robot was moving its arm. An embodied physical robot would be able to use objects in its lessons. It would be able to recognize and hold objects, and thus to engage in lessons in which the focus lies not on the materials (as was the case in our study, due to the tablet) but on the robot and the child interacting with these materials.

Lesson 4: Explore

Lesson four is to explore. Infants learn by engaging in actions with no apparent goal. Such actions help them to learn, amongst others, about action-consequence sequences and about the affordances of objects in a particular spatial lay-out, and also about less obvious affordances for action leading to new uses of objects (e.g., as in discovering tools).Children's exploration can be regarded as very rapidly learning in real-time about objects and what they afford in a given situation, which underlies adaptivity and creativity (Oudgenoeg-Paz et al., 2016). Our current robot cannot respond to events which are not pre-programed in the script, and cannot change its lesson and instruction behavior depending on new, not pre-programmed events or object structures in the environment. Exploration in the sense of rapid real-time learning of action possibilities may be necessary for a robot to become truly adaptive. It can perceive the environment, draw the learner's attention to relevant or new stimuli in the environment, and respond meaningfully to unexpected events.

Lesson 5: Be social

The fifth lesson is to be social and this may be the most difficult challenge. Infants learn social behavior through imitating their parents, and the parents provide social information (such as facial expressions and vocalizations) which the infant can imitate, matching the infant's developmental stage. However, it is not merely about imitation, or 'echoing', social and emotional cues expressed by the face, body posture, movement patterns of others (a challenge which could be in principle technically mastered by robots in due time). It is also about the direct coupling of this echoing, mimicking and imitation of others' behavior to the child's own emotion systems (Gallese & Cuccio, 2015), enabling what philosophers of mind call direct access to the feeling states of others, underlying empathy and sympathy, giving motivational power to social (rule-following) and moral behavior (Tomasello & Vaish, 2013). An embodied intelligent agent can adapt its social behavior to the child’s needs. It also can couple action and sounds, but this can only be done manually – that is, through human interpretation and empathy - in the robot used in this study. Likewise, the robot’s gestures in our study had to be time-locked to its speech by us, and thus may have differed from how humans would combine language and gestures naturally.

Lesson 6: Learn a language

The sixth and last lesson is to learn a language, which is another challenging task. Language is a symbol system, in which sounds are arbitrarily mapped onto meaning, while combinatorial rules (morphology and syntax) specify how smaller elements can be combined into larger units such as words and sentences. Language-in-use is also a system to share meaning in communicative interactions through arbitrary but usually well-understood symbols that refer to the real world. Language as a symbol system can be abstracted from the real world, disregarding the referential meaning of language. Language, in this sense, is a computational system of word-word relations, but its connection to the real world state of affairs, actions and events is problematic (Pulvermüller, 2013). Robots place us for a challenging question: What is *true* language comprehension and what is *true* communicative use of language? The current robot can speak, but cannot be said to have any comprehension of its utterances in terms of word-world relations. The sounds it produces are, referentially, as meaningless to the robot as any other sound. The robot can detect sound and convert speech streams from adults into text which it can subsequently use to respond (that is, reacting to the occurrence of particular key words), but it still does not have any comprehension of the adult’s speech.

Recognition of children’s speech is still a hurdle hard to take. Although it can be expected that this hurdle can be overcome in due time, this will still beg the question if the robot indeed understands what a child is saying. Some natural-language processing and generation systems have been developed much further and can receive and produce speech without developers scripting each and every answer beforehand. However, do such systems truly have language? They do not have their concepts grounded in physical interactions with the environment nor in empathy-based social interactions with others, and perhaps such interactions are necessary to truly comprehend and use language with all its subtle meanings. An embodied intelligent robot agent that would have similar concepts as its language would engage in very different interactions than our current robot. For example, it could use child-directed speech, interpret the child's current understanding and intentions, and use its knowledge grounded in interactions with the environment to gesture and act-out.

Discussion

These six lessons illustrate that many technological developments are needed before it would be possible to develop an embodied intelligent robot agent that could deploy a robot's full potential in educational situations. Some of these technological developments are already nearby, others will take more time. Some other requirements may be impossible to meet. Apart from the question whether it is possible to develop robots in such a way, however, the question arises whether it is *desirable* to develop robots in such a way. In the most optimistic scenario, such an embodied intelligent robot agent would be capable of imitating human teachers and likely to be a very effective teacher. However, this is only true if children actually respond to a robot tutor the way they respond to a human tutor. This is still a relatively unexplored area of research.

In a recent study, we have taken a first step in this direction with a study on children following a robot's *versus* a human's non-verbal behavior (Verhagen, van den Berghe, Oudgenoeg-Paz, Küntay, & Leseman, in press). An implicit assumption in the field of human-robot interaction is that a robot's non-verbal behavior, such as eye gazing and pointing, is picked up by children and interpreted as referential cue, much like in adult-child interactions. To test this assumption, we compared children's reliance on a robot's versus a human's eye gaze and pointing behaviors to determine what the robot or human referred to. Children were presented with two pictures and a referential conflict: A robot or an adult pointed or gazed at one picture but verbally labeled the other. Children's resolution of this conflict shows the weight they attribute to the non-verbal behavior as opposed to the verbal information. Results showed that children did not rely differently on the non-verbal cues of a robot than on those of a human. Overall, they relied much more on the pointing cue than the gaze cue in line with the research literature on referential communication involving human interaction partners (Hansen & Markman, 2009; Grassmann & Tomasello, 2010; Verhagen, Grassmann, & Küntay, 2017).

Interestingly, children's perception of the robot also played a role. Children who perceived the robot as more human-like relied on pointing more strongly when presented with a novel label versus a familiar label, whereas children who perceived the robot as less human-like did not show this difference. Implications of the findings entail that non-verbal behaviors such as eye gaze and pointing indeed can be used in child-robot interactions to support children's learning. However, children's perception of the robot as being humanoid may be crucial for establishing trustworthiness. Our study on children's anthropomorphism of robots (described in Chapter 6) showed that many children have a tendency to attribute human-like characteristics to robots, such as having mental states and high-functioning cognitive abilities, but we also found indications that these high expectations risk to be frustrated and adversely affect learning with the robot, due to the technical limitations. Nonetheless, together these two studies show that at least some children appear to treat robots very similarly to humans.

This raises fundamental ethical questions. Given that children seem to perceive robots more as similar to humans the more robots display human-like features, is it ethical to develop robots that are even more similar to humans? Robots do not have the ability to truly understand feelings, nor do they have a moral compass or empathy-based motivation to care for children, while their highly human-like behavior may lead people, especially young children, to believe that they do. In a very basic sense we are deliberately deceiving the children. The ethics of developing robots for education should, therefore, be given a much more central place in the field (Smakman, 2018).

Implications and Recommendations

Taking all this together, it is doubtful whether it will be possible to develop a robot according to the six lessons of Smith and Gasser (2005), that is, to develop a robot that can be multimodal, incremental, physical, social, explore, and master language. Without concluding that is not possible, it seems certain that the required technological developments demand huge investments and the question is whether these investments will ultimately pay-off. Perhaps it is more worthwhile to take a different approach to developing robots for education. Instead of trying to develop robots that can copy human tutors, we should look for ways in which robots can *complement* humans in education. If a robot is designed to copy a human, it is likely that it will inevitably fall short of children's expectations sooner or later, at least given the current state of technology, and it will raise serious ethical questions. Robots simply cannot behave exactly like humans. Rather, we should look how the different type of intelligence of robots can be used in an optimal way. For example, compared to humans, robots have infinite patience, do not get bored, can be designed specifically to serve, have a virtually unlimited memory capacity (through connections to cloud services), have computational power, and a potentially unlimited repository of knowledge (e.g., through connections with the internet). And, indeed, robots can ‘speak’ and ‘recognize’ multiple languages, carry multiple languages’ grammars, dictionaries and stories, as was one of the current project’s starting points. Such qualities are especially valuable for simple tasks in which the learner needs extensive practice but does not need the robot to have deep understanding and highly interactional qualities. For example, a robot could help children learn the tables and solve mathematical equations, or supervise independent seatwork or motivate children to do their homework. And, indeed, a robot can help language learning children by providing them with the dictionary items or grammatical examples needed in particular ‘simple’ language learning tasks, like learning an L2 vocabulary or translating words. In such tasks, the robot’s function is clear and does not mislead children by appearing much more communicative and socially skilled than it actually is. There are many situations in which robots can have a contribution to education, and in which we can clearly manage children’s expectations beforehand to make sure that the robot will not disappoint children.

Another area to pursue in incorporating robots in education is related to the ethical question we raised. Given that robots, also intelligent robots, will enter our daily work and private life in the near future, we need new educational programs that focus on teaching children to ‘understand’ robots. Children can be taught how to interact with robots, what robots can and cannot do, what impression they may evoke, and to what extent such impressions are true or false, and so forth. Robots are, in many respects, a new species that we still do not understand well, unlike other nonhuman species that populate our (domestic) environments since ages. Equipping children with such knowledge about technologies as robotics and artificial intelligence will enable them to function in a more conscious and critical manner in a world where these technologies are increasingly incorporated in daily life.

General Conclusion

This dissertation has shown that despite the potential of robots for language education, there are many technological limitations that have prevented us from designing robot-assisted lessons that have added value over other forms of technology for L2 learning. Several lessons can be drawn for future design of RALL interactions. The first is that a tablet in itself does not hamper learning as compared to physical objects, but that an interaction that revolves around the tablet too much may cancel out any possible benefits of the robot. The second lesson is that a robot's verbal and non-verbal behavior should be designed with care to ensure that the robot benefits learning without distracting children. The last is that differences between children in the perception of the robot and their language and attention skills affect their interaction with the robot and their learning, and should be taken into account when designing RALL interactions.

There are some limitations to the research reported in this dissertation. The first is that the robot in the L2TOR system could not realize the many advantages that robots potentially have over other forms of technology, discussed in the Introduction and Chapter 2. Thus, the robot used in this dissertation is far from the optimal (robot) language tutor. Moreover, we did not compare the robot to other forms of technology in this dissertation. We could not design RALL interactions that did not include a mediating device, due to the technological limitations of the robot. We could only compare the L2TOR setup consisting of a robot and a tablet to a tablet only, or to playing language games on a tablet with a child peer. Therefore, we cannot draw firm conclusions regarding the effectiveness of robot as compared to other types of technology. Despite these limitations, this dissertation also made a contribution to RALL research by critically evaluating current research in RALL, addressing issues in designing RALL interactions, and by including the first large-scale randomized controlled trial in RALL research.

In the future, robots should be developed in a way which makes them much more embodied than they are now, following the six lessons from Smith and Gasser (2005). However, certain limitations seem impossible to overcome. Moreover, rather than trying to make robots as similar to humans as possible, perhaps we should focus on investigating, while preserving the advantages over other technology, whether there are ways in which robots can *complement* humans.

References

References marked with an asterisk indicate studies that were included in the review presented in Chapter 2.

* Alemi, M., Meghdari, A., & Ghazisaedy, M. (2014). Employing humanoid robots for teaching English language in Iranian junior high-schools. *International Journal of Humanoid Robotics, 11,* 1450022-1-1450022-25.

* Alemi, M., Meghdari, A., & Ghazisaedy, M. (2015). The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition. *International Journal of Social Robotics, 7,* 523–535.

Alemi, M., Meghdari, A., Mahboub Basiri, N., & Taheri, A. (2015b). The effect of applying humanoid robots as teacher assistants to help Iranian autistic pupils learn English as a foreign language. In *Proceedings of the 7th International Conference on Social Robotics* (pp. 1–10).

* Alemi, M., Meghdari, A., & Sadat Haeri, N. (2017). Young EFL learners' attitude towards RALL: An observational study focusing on motivation, anxiety, and interaction. In *Proceedings of the International Conference on Social Robotics* (pp. 252–261).

Antonucci, S. M., & Alt, M. (2011). A lifespan perspective on semantic processing of concrete concepts: Does a sensory/motor model have the potential to bridge the gap? *Cognitive, Affective, & Behavioral Neuroscience*, *11*, 551–572.

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology, 8,* 19–32.

Arsénio, A. M. (2014). Developmental language learning from human/humanoid robot social interactions: An embodied and situated approach. In *Robotics: Concepts, Methodologies, Tools, and Applications* (pp. 1328–1353). Hershey, PA: IGI Global.

Axelsson, E. L., Williams, S. E., & Horst, J. S. (2016). The effect of sleep on children's word retention and generalization. *Frontiers in Psychology*, *7*.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158–173.

Baguley, T. S. (2012). Serious stats: A guide to advanced statistics for the behavioral sciences. New York: Palgrave Macmillan

Barcelona European Council, 15 and 16 March 2002, *Presidency Conclusions, part I, 43.1*.

Barrett, N. E., & Liu, G.-Z. (2016). Global trends and research aims for English academic oral presentations: Changes, challenges, and opportunities for learning technology. *Review of Educational Research, 86,* 1227–1271.

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, *59*(1), 617–645.

Bartneck, C., & Forlizzi, J. (2004). A design-centred framework for social human-robot interaction. In *Proceedings of the Ro-Man 2004* (pp. 591–594).

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*(1), 71-81.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Beck, I. L., & McKeown, M. G. (1985). Teaching vocabulary: Making the instruction fit the goal. *Educational Perspectives*, *23*(1), 11-15.

Belpaeme, T., Kennedy, J., Baxter, P., Vogt, P., Krahmer, E. E., Kopp, S., ... & Pandey, A. K. (2015). L2TOR-second language tutoring using social robots. In *Proceedings of the ICSR 2015 WONDER Workshop*.

Belpaeme, T., Vogt, P., Van den Berghe, R., Bergmann, K., Göksun, T., De Haas, M., ... & Papadopoulos, F. (2018). Guidelines for designing social robots as second language tutors. *International Journal of Social Robotics*, 1-17.

Beran, T. N., Ramirez-Serrano, A., Kuzyk, R., Fior, M., & Nugent, S. (2011). Understanding how children understand robots: Perceived animism in child–robot interaction. *International Journal of Human-Computer Studies*, *69*(7-8), 539-550.

Bernstein, D., & Crowley, K. (2008). Searching for signs of intelligent life: An investigation of young children's beliefs about robot intelligence. *The Journal of the Learning Sciences*, *17*(2), 225-247.

Blom, E. (2019). What everyone should know about language development in children [Oration].

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2015). A quasi-universal nonword repetition task as a diagnostic tool for bilingual children learning Dutch as a second language. *Journal of Speech, Language, and Hearing Research*, *58*(6), 1747-1760.

Breazeal, C., Kidd, C. D., Thomaz, A. L., Hoffman, G., & Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 708-713).

Brenders, P., van Hell, J. G., & Dijkstra, T. (2011). Word recognition in child second language learners: Evidence from cognates and false friends. *Journal of Experimental Child Psychology*, *109*(4), 383-396.

Burnett, C. (2010). Technology and literacy in early childhood educational settings: A review of research. *Journal of Early Childhood Literacy*, *10*(3), 247–270.

Caporael, L. R. (1986). Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in Human Behavior*, *2*(3), 215-234.

Carlo, M.S., McLaughlin, B., Snow, C.E., Dressler, C., Lippman, D.N., Lively, T.J., & White, C.E. (2004). Closing the gap: Addressing the vocabulary needs of English language learners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39*(2), 188-215.

Chandra, S., Paradeda, R., Yin, H., Dillenbourg, P., Prada, R., & Paiva, A. (2018). Do children perceive whether a robotic peer is learning or not?. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 41-49).

Cheung, H. (1996). Nonword span as a unique predictor of second-language vocabulary language. *Developmental Psychology*, *32*(5), 867-873.

Chiat, S. (2015). Non-word repetition. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Methods for assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 227–250). Bristol: Multilingualism Matters.

Chin, K., Hong, Z.-W., & Chen, Y.-L. (2014). Impact of using an educational robot-based learning system on students' motivation in elementary education. *IEEE Transactions on Learning Technologies, 7,* 335-345.

Collins, M. F. (2010). ELL preschoolers' English vocabulary acquisition from storybook reading. *Early Childhood Research Quarterly*, *25*(1), 84-97.

Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educational Psychology Review*, *26*(2), 197-223.

Creese, A., & Blackledge, A. (2010). Translanguaging in the bilingual classroom: A pedagogy for learning and teaching?. *The Modern Language Journal*, *94*(1), 103-115.

Cronin, P., Ryan, F., & Coughlan, M. (2008). Undertaking a literature review: A step-by-step approach. *British Journal of Nursing,* 17, 38–43.

Cummins, J. (1991). Interdependence of first-and second-language proficiency in bilingual children. *Language Processing in Bilingual Children*, 70-89.

Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In: Street, B. & Hornberger, N. H. (Eds.). *Encyclopedia of Language and Education*, *2nd Edition, Volume 2: Literacy.* (pp. 71-83). New York: Springer Science.

Dautenhahn, K., & Werry, I. (2004). Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmatics & Cognition*, *12*(1), 1-35.

* de Haas, M., Baxter, P., de Jong, C., Krahmer, E., & Vogt, P. (2017). 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* (pp. 127–128).

de Nooijer, J. A., van Gog, T., Paas, F., & Zwaan, R. A. (2013). Effects of imitating gestures during encoding or during retrieval of novel verbs on children's test performance. *Acta Psychologica*, *144*(1), 173-179.

* de Wit, J., Schodde, T., Willemsen, B., Bergmann, K., de Haas, M., Kopp, S., … 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* (pp. 50–58).

Diekelmann, S., Wilhelm, I., & Born, J. (2009). The whats and whens of sleep-dependent memory consolidation. *Sleep Medicine Reviews*, *13*, 309–321.

Dörnyei, Z. (1994). Motivation and motivating in the foreign language classroom. *The Modern Language Journal, 78,* 273–284.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3-4), 177-190.

Dunn, L. M., Dunn, L. M., & Schlichting, L. (2005). *Peabody picture vocabulary test-III-NL*. Amsterdam: Pearson.

* Eimler, S., von der Pütten, A., & Schächtle, U. (2010). Following the white rabbit – A robot rabbit as vocabulary trainer for beginners of English. In *Proceedings of the 6th Symposium of the Workgroup Human Computer Interaction and Usability Engineering* (pp. 322–339).

Ellis, N., & Beaton, A. (1993). Factors affecting the learning of foreign language vocabulary: Imagery keyword mediators and phonological short-term memory. *The Quarterly Journal of Experimental Psychology*, *46*(3), 533-558.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864-886.

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, *26*(2), 143-155.

Esser, H. (2006). *Migration, language and integration.* AKI Research Review 4. Berlin: Programme on Intercultural Conflicts and Societal Integration (AKI), Social Science Research Center.

Eyssel, F., Kuchenbrandt, D., Hegel, F., & de Ruiter, L. (2012). Activating elicited agent knowledge: How robot and user features shape the perception of social robots. In *Proceedings of the 2012 RO-MAN* (pp. 851-857).

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal, 22,* 338–342.

Fernaeus, Y., Håkansson, M., Jacobsson, M., & Ljungblad, S. (2010). How do you play with a robotic toy animal? A long-term study of Pleo. In *Proceedings of the 9th International Conference on Interaction Design and Children* (pp. 39–48).

Fink, J. (2012). Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *International Conference on Social Robotics* (pp. 199-208).

Funakoshi, K., Mizumoto, T., Nagata, R., & Nakano, M. (2011). The chanty bear: A newapplication for HRI research. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (pp. 141–142).

Gallese, V., & Cuccio, V. (2015). The paradigmatic body: Embodied simulation, intersubjectivity, the bodily self, and language. In T. Metzinger & J.M. Windt (Eds.), *Open Mind*: 14(T) (pp. 1-22). Frankfurt am Main: Open Mind Group.

Gallese, V., & Lakoff, G. (2005). The brain’s concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, *22*(3–4), 455–479.

Garrett, N. (1991). Technology in the service of language learning: Trends and issues. *The Modern Language Journal, 75,* 74-101.

Garrett, N. (2009). Computer-assisted language learning trends and issues revisited: Integration innovation. *The Modern Language Journal, 93,* 719–740.

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*, 513–543.

Gathercole, S. E., & Baddeley, A. D. (1990). The role of phonological memory in vocabulary acquisition: A study of young children learning new names. *British Journal of Psychology*, *81*, 439–454.

Gibson, E. J., & Pick, A. D. (2000). *An ecological approach to perceptual learning and development*. New York, NY: Oxford University Press.

Gibson, J. J. (1986). *The ecological approach to visual perception*. Hillsdale, NJ: Erlbaum.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, *20*, 1–55.

Glenberg, A. M., Brown, M., & Levin, J. R. (2007). Enhancing comprehension in small reading groups using a manipulation strategy. *Contemporary Educational Psychology*, *32*(3), 389–399.

Glenberg, A. M., Goldberg, A. B., & Zhu, X. (2011). Improving early reading comprehension using embodied CAI. *Instructional Science*, *39*(1), 27–39.

Glenberg, A. M., Gutierrez, T., Levin, J. R., Japuntich, S., & Kaschak, M. P. (2004). Activity and imagined activity can enhance young children's reading comprehension. *Journal of Educational Psychology*, *96*(3), 424–436.

Glenberg, A. M., Willford, J., Gibson, B., Goldberg, A., & Zhu, X. (2011). Improving reading to improve math improving reading to improve math. *Scientific Studies of Reading*, 1–25.

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition*, *35*(3), 483-517.

Godfroid, A., & Schmidtke, J. (2013). What do eye movements tell us about awareness? A triangulation of eye-movement data, verbal reports and vocabulary learning scores. *Noticing and Second Language Acquisition: Studies in Honor of Richard Schmidt*, 183-205.

Gogate, L. J., & Hollich, G. (2010). Invariance detection within an interactive system: A perceptual gateway to language development. *Psychological Review*, *117*(2), 496-516.

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. F. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, *27*(1), 70–105.

Gomez, E. A., Wu, D., & Passerini, K. (2010). Computer-supported team-based learning: The impact of motivation, enjoyment and team contributions on learning outcomes. *Computers & Education*, *55*(1), 378-390.

* Gordon, G., Breazeal, C., & Engel, S. (2015). Can children catch curiosity from a social robot? In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 91–98).

* Gordon, G., Spaulding, S., Kory Westlund, J. M., Lee, J. J., Plummer, L., Martinez, M., … Breazeal, C. L. (2016). Affective personalization of a social robot tutor for children's second language skills. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (pp. 3951–3957).

Grassmann, S., & Tomasello, M. (2010). Young children follow pointing over words in interpreting acts of reference. *Developmental Science, 13,* 252-263.

Han, J., Kang, B., Park, S., & Hong, S. (2012). How to sustain long-term interaction between children and ROBOSEM in English class. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 421–422).

* Han, J.-H., Jo, M.-H., Jones, V., & Jo, J.-H. (2008). Comparative study on the educational use of home robots for children. *Journal of Information Processing Systems, 4,* 159-168.

Hansen, M.B., & Markman, E.M., (2009). Children's use of mutual exclusivity to learn labels for parts of objects. *Developmental Psychology, 45,* 592-596.

Hegel, F., Krach, S., Kircher, T., Wrede, B., & Sagerer, G. (2008). Understanding social robots: A user study on anthropomorphism. In *The 17th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 574-579).

Heitink, M., Fisser, P., & Voogt, J. (2013). Learning vocabulary through a serious game in Primary Education. In *Proceedings of Society for Information Technology & Teacher Education International Conference 2013* (pp. 2845–2850).

Hemsley, G., Holm, A., & Dodd, B. (2013). Conceptual distance and word learning: patterns of acquisition in Samoan–English bilingual children. *Journal of Child Language*, *40*(4), 799-820.

* Herberg, J. S., Feller, S., Yengin, I., & Saerbeck, M. (2015). Robot watchfulness hinders learning performance. In *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 153–160).

Hockema, S. A., & Smith, L. B. (2009). Learning your language, outside-in and inside-out. *Linguistics*, *47*(2), 453–479.

Hoff, E. (2013). Interpreting the early language trajectories of children from low SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology, 49*(1), 4–14.

* Hong, Z.-W., Huang, Y.-M., Hsu, M., & Shen, W.-W. (2016). Authoring robot-assisted instructional materials for improving learning performance and motivation in EFL classrooms. *Educational Technology & Society, 19,* 337–349.

Hood, D., Lemaignan, S., & Dillenbourg, P. (2015). When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 83–90).

Hornberger, N. H. (2005). Opening and filling up implementational and ideological spaces in heritage language education. *The Modern Language Journal*, *89*(4), 605-609.

Howell, S. R., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language*, *53*, 258–276.

* Hsiao, H.-S., Chang, C.-S., Lin, C.-Y., & Hsu, H.-L. (2012). "iRobiQ ": The influence of bidirectional interaction on kindergarteners' reading motivation, literacy, and behavior. *Interactive Learning Environments, 23,* 269–292.

* Hyun, E., Kim, S., Jang, S., & Park, S. (2008). Comparative study of effects of language instruction program using intelligence robot and multimedia on linguistic ability of young children. In *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 187–192).

* In, J.-Y., & Han, J.-H. (2015). The acoustic-phonetic change of English learners in robot assisted learning. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 39–40).

Iverson, J. M. (2010). Developing language in a developing body: The relationship between motor development and language development. *Journal of Child Language, 37,* 229-261.

Jalongo, M. R., & Sobolak, M. J. (2011). Supporting young children's vocabulary growth: The challenges, the benefits, and evidence-based strategies. *Early Childhood Education Journal*, *38*(6), 421-429.

James, K. H., & Bose, P. (2014). Self-generated actions during learning objects and sounds create sensori-motor systems in the developig brain. *Cognition, Brain, Behavior: An Interdisciplinary Journal*, *15*(4), 485–203.

James, K. H., & Swain, S. N. (2011). Only self-generated actions create sensori-motor systems in the developing brain. *Developmental Science*, *14*(4), 673–678.

Jiang, N. (2004). Semantic transfer and its implications for vocabulary teaching in a second language. *The Modern Language Journal, 88,* 417-432.

Jipson, J. L., & Gelman, S. A. (2007). Robots and rodents: Children's inferences about living and nonliving kinds. *Child Development*, *78*(6), 1675-1688.

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., ... & Shen, S. (2012). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology*, *48*(2), 303-314.

* Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction, 19,* 61–84.

Kanero, J., Geçkin, V., Oranç, C., Mamus, E., Küntay, A. C., & Göksun, T. (2018). Social robots for early language learning: Current evidence and future directions. *Child Development Perspectives, 12,* 146–151. doi:10.1111/cdep.12277

Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, *24*(2), 313-334.

Kennedy, J., Baxter, P., & Belpaeme, T. (2015a). Comparing robot embodiments in a guided discovery learning interaction with children. *International Journal of Social Robotics*, *7*(2), 293-308.

Kennedy, J., Baxter, P., & Belpaeme, T. (2015b). 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* (pp. 67-74).

Kennedy, J., Baxter, P., Senft, E., & Belpaeme, T. (2015). Higher nonverbal immediacy leads to greater learning gains in child-robot tutoring interactions. In *Proceedings of the International Conference on Social Robotics* (pp. 327-336).

* Kennedy, J., Baxter, P., Senft, E., & Belpaeme, T. (2016). Social robot tutoring for child second language learning. In *Proceedings of the Eleventh ACM/IEEE International Conference on Human Robot Interaction* (pp. 231–238).

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., … Belpaeme, T. (2017). Child speech recognition in human-robot interaction: Evaluations and recommendations. In *Proceedings of the 2017 ACM/IEEE Human Robot Interaction Conference* (pp. 82–90).

Kersten, A. W., & Smith, L. B. (2002). Attention to novel objects during verb learning. *Child Development, 73,* 93–109.

Kidd, C. D. (2003). *Sociable robots: The role of presence and task in human-robot interaction* (Doctoral dissertation, Massachusetts Institute of Technology).

Kidd, C. D., & Breazeal, C. (2004). Effect of a robot on user perceptions. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 3559-3564).

King, A., Staffieri, A., & Adelgais, A. (1998). Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *Journal of Educational Psychology*, *90*(1), 134.

Kita, S., & Özyürek, A. (2007). *How does spoken language shape iconic gestures.* In: Duncan, S., Cassell, J., Levy, E.T. (Eds.), Gesture and the Dynamic Dimension of Language: Essays in Honor of David McNeill. John Benjamins Publishing Company, Amsterdam, pp. 67-74.

Kory Westlund, J., & Breazeal, C. (2015). The interplay of robot language level with children's language learning during storytelling. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 65–66).

* Kory Westlund, J., Dickens, L., Jeong, S., Harris, P., DeSteno, D., & Breazeal, C. (2015). A comparison of children learning new words from robots, tablets, & people. In *Conference Proceedings of New Friends: The 1st International Conference on Social Robots in Therapy and Education* (pp. 7–8).

* Kory Westlund, J. M., Dickens, L., Jeong, S., Harris, P. L., DeSteno, D., & Breazeal, C. L. (2017). Children use non-verbal cues to learn new words from robots as well as people. *International Journal of Child-Computer Interaction, 13,* 1–9.

* Kory Westlund, J. M., Jeong, S., Park, H. W., Ronfard, S., Adhikari, A., Harris, P. L., … Breazeal, C. L. (2017). Flat vs. expressive storytelling: Young children's learning and retention of a social robot's narrative. *Frontiers in Human Neuroscience, 11.*

Kory Westlund, J. M. K., Martinez, M., Archie, M., Das, M., & Breazeal, C. (2016). Effects of framing a robot as a social agent or as a machine on children's social behavior. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 688-693).

Köse, H., Uluer, P., Akalın, N., Yorgancı, R., Özkul, A., & Ince, G. (2015). The effect of embodiment in sign language tutoring with assistive humanoid robots. *International Journal of Social Robotics, 7,* 537–548.

Lan, Y.-J., Fang, S., Legault, J., & Li, P. (2015). Second language acquisition of Mandarin Chinese vocabulary: context of learning effects. *Educational Technology Research and Development*, *63*, 671–690.

Lee, E., & Lee, Y. (2008). A pilot study of intelligent robot aided education. In *Proceedings of the ICCE Conference on Advanced Learning Technologies, Open Contents, & Standards* (pp. 595–596).

* Lee, S., Noh, H., Lee, J., Lee, K., Lee, G. G., Sagong, S., & Kim, M. (2011). On the effectiveness of robot-assisted language learning. *ReCALL, 23,* 25–58.

Leeuwestein, H. Oudgenoeg-Paz, O., Verhagen, J., Vogt, P., Spit, S., Barking, M., … Leseman, P. (in preparation). Teaching Turkish-Dutch kindergartners Dutch vocabulary with a social robot: Does the robot's use of Turkish translations benefit children's Dutch vocabulary learning?.

Leite, I., Martinho, C., & Paiva, A. (2013). Social robots for long-term interaction: A survey. *International Journal of Social Robotics, 5,* 291–308.

Leseman, P.P.M., Henrichs, L.F., Blom, E. & Verhagen, J. (2019). Young mono- and bilingual children’s exposure to academic language as related to language development and school achievement. In V. Grøver, P. Ucelli, M. Rowe & E. Lieven (Eds), *Learning through language* (pp. 205-217). Cambridge, MA: Cambridge University Press.

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*, *109*(5), 1431-1436.

Leyzberg, D., Spaulding, S., & Scassellati, B. (2014). Personalizing robot tutors to individuals' learning differences. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 423-430).

Leyzberg, D., Spaulding, S., Toneva, M., & Scassellati, B. (2012). The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 1882–1887).

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning, 60,* 309–365.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: The Guilford Press.

Liu, E. Z. F., Lin, C. H., & Chang, C. S. (2010). Student satisfaction and self-efficacy in a cooperative robotics course. *Social Behavior and Personality: An International Journal, 38,* 1135–1146.

Liu, S., Liao, H., & Pratt, J. A. (2009). Impact of media richness and flow on e-learning technology acceptance. *Computers & Education, 52,* 599–607.

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

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

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children’s word learning: A meta-analysis. *Review of Educational Research*, *80*(3), 300-335.

Masoura, E. V., & Gathercole, S. E. (2005). Contrasting contributions of phonological short-term memory and long-term knowledge to vocabulary learning in a foreign language. *Memory*, *13*(3/4), 422–429.

Mavilidi, M., Okely, A. D., Chandler, P., Cliff, D. P., & Paas, F. (2015). Effects of integrated physical exercises and gestures on preschool children's foreign language vocabulary learning. *Educational Psychology Review*, *27*, 413–426.

* Mazzoni, E., & Benvenuti, M. (2015). A robot-partner for preschool children learning English using socio-cognitive conflict. *Educational Technology & Society, 18,* 474-485.

McClelland, M. M., Cameron, C. E., Wanless, S. B., & Murray, A. (2007). Executive function, behavioral self-regulation, and social-emotional competence. *Contemporary perspectives on social learning in early childhood education*, *1*, 113-137.

McGonigle, B., & Chalmers, M. (2001). Spatial representations as cause and effect: Circular causality comes to cognition. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (pp. 247–277). Cambridge, MA, USA: The MIT Press.

* Meiirbekov, S., Balkibekov, K., Jalankuzov, Z., & Sandygulova, A. (2016). "You win, I lose": Towards adapting robot's teaching strategy. In *Proceedings of the Eleventh ACM/IEEE International Conference on Human Robot Interaction* (pp. 475–476).

Mercer, N., & Littleton, K. (2007). Dialogue and the development of children's thinking: A sociocultural approach. London: Routledge.

Metsala, J. L. (1999). Young children's phonological awareness and nonword repetition as a function of vocabulary development. *Journal of Educational Psychology*, *91*(1), 3.

Miller, E. (2005). Fighting technology for toddlers. *Education Digest*, *71*(3), 55–58.

Mitnik, R., Recabarren, M., Nussbaum, M., & Soto, A. (2009). Collaborative robotic instruction: A graph teaching experience. *Computers & Education, 53,* 330–342.

Monaco, C., Mich, O., Ceol, T., & Potrich, A. (2018). Investigating Mental Representations about Robots in Preschool Children. *arXiv preprint arXiv:1806.03248*.

Mondria, J.-A., & Wiersma, B. (2004). Receptive, productive, and receptive + productive L2 vocabulary learning: What difference does it make? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a Second Language* (pp. 79–100).

Moon, A., Troniak, D. M., Gleeson, B., Pan, M. K., Zheng, M., Blumer, B. A., ... & Croft, E. A. (2014). Meet me where I'm gazing: How shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 334-341).

Moriguchi, Y., Kanda, T., Ishiguro, H., Shimada, Y., & Itakura, S. (2011). Can young children learn words from a robot? *Interaction Studies, 12,* 107–118.

Moriguchi, Y., Okanda, M., & Itakura, S. (2008). Young children's yes bias: How does it relate to verbal ability, inhibitory control, and theory of mind?. *First Language*, *28*(4), 431-442.

* Movellan, J. R., Eckhardt, M., Virnes, M., & Rodriguez, A. (2009). Sociable robot improves toddler vocabulary skills. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (pp. 307–308).

Mulder, H., Hoofs, H., Verhagen, J., van der Veen, I., & Leseman, P. P. M. (2014). Psychometric properties and convergent and predictive validity of an executive function test battery for two-year-olds. *Frontiers in Psychology*, *5*.

Muthen, L. K., & Muthén, B. O. (2010). *Mplus user's guide, v. 6.1.* Los Angeles, CA: Muthén & Muthén.

Myung, J., Blumstein, S. E., & Sedivy, J. C. (2006). Playing on the typewriter, typing on the piano: Manipulation knowledge of objects. *Cognition*, *98*, 223–243.

Nagata, R., Mizumoto, T., Funakoshi, K., & Nakano, M. (2010). Toward a chanting robot for interactively teaching English to children. In *Proceedings of INTERSPEECH Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology* (pp. 2–13).

Nikolov, M., & Djigunović, J. M. (2006). Recent research on age, second language acquisition, and early foreign language learning. *Annual Review of Applied Linguistics*, *26*, 234-260.

Nourbakhsh, I. R., Crowley, K., Bhave, A., Hamner, E., Hsiu, T., Perez-Bergquist, A., … Wilkinson, K. (2005). The robotic autonomy mobile robotics course: Robot design, curriculum. *Autonomous Robots, 18,* 103–127.

O'Donnell, A. M. (1999). Structuring dyadic interaction through scripted cooperation. *Cognitive Perspectives on Peer Learning*, 179-196.

O'Neill, D. K., Topolovec, J., & Stern-Cavalcante, W. (2002). Feeling sponginess: The importance of descriptive gestures in 2- and 3-year-old children's acquisition of adjectives. *Journal of Cognition and Development*, *3*(3), 243–277.

Obaid, M., Barendregt, W., Alves-Oliveira, P., Paiva, A., & Fjeld, M. (2015). Designing robotic teaching assistants: interaction design students' and children's views. In *International Conference on Social Robotics* (pp. 502-511).

Okita, S. Y., Ng-Thow-Hing, V., & Sarvadevabhatla, R. (2009). Learning together: ASIMO developing an interactive learning partnership with children. In *The 18th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 1125-1130).

Öttl, B., Dudschig, C., & Kaup, B. (2017). Forming associations between language and sensorimotor traces during novel word learning. *Language and Cognition*, *9*, 156-171.

Oudgenoeg-Paz, O., Leseman, P. P., & Volman, M. (2015). Exploration as a mediator of the relation between the attainment of motor milestones and the development of spatial cognition and spatial language. *Developmental Psychology*, *51*(9), 1241-1253.

Oudgenoeg-Paz, O., Volman, M. J. M., & Leseman, P. P. M. (2016). First steps into language? Examining the specific longitudinal relations between walking, exploration and linguistic skills. *Frontiers in Psychology*, *7*.

Papagno, C., Valentine, T., & Baddeley, A. (1991). Phonological short-term memory and foreign-language vocabulary learning. *Journal of Memory and Language*, *30*, 331-347.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist, 37,* 91–105.

Penno, J. F., Wilkinson, I. A., & Moore, D. W. (2002). Vocabulary acquisition from teacher explanation and repeated listening to stories: Do they overcome the Matthew effect?. *Journal of Educational Psychology*, *94*(1), 23-33.

Pereira, A., Martinho, C., Leite, I., & Paiva, A. (2008). iCat, the chess player: The influence of embodiment in the enjoyment of a game. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 3* (pp. 1253-1256).

Pulvermüller, F. (2013). Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits*. Brain & Language, 127*, 86-103.

Rasku-Puttonen, H., Lerkkanen, M. K., Poikkeus, A. M., & Siekkinen, M. (2012). Dialogical patterns of interaction in pre-school classrooms. *International Journal of Educational Research*, *53*, 138-149.

R Core team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.r-project.org/.

Riek, L. D., Rabinowitch, T. C., Chakrabarti, B., & Robinson, P. (2009). How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (pp. 245-246).

* Rintjema, E., van den Berghe, R., Kessels, A., de Wit, J., & Vogt, P. (2018). A robot teaching young children a second language: The effect of multiple interactions on engagement and performance. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 219–220).

Rispens, J., & Baker, A. (2012). Nonword repetition: The relative contributions of phonological short-term memory and phonological representations in children with language and reading impairment. *Journal of Speech, Language, and Hearing Research*, *55*(3), 683-694.

Robinson, P. (1995). Attention, memory, and the "noticing" hypothesis. *Language Learning*, *45*(2), 283–331.

Rohrbeck, C. A., Ginsburg-Block, M. D., Fantuzzo, J. W., & Miller, T. R. (2003). Peer-assisted learning interventions with elementary school students: A meta-analytic review. *Journal of Educational Psychology, 95,* 240–257.

* Rosenthal-von der Pütten, A. M., Straßmann, C., & Krämer, N. C. (2016). Robots or agents-neither helps you more or less during second language acquisition. Experimental study on the effects of embodiment and type of speech output on evaluation and alignment. In *Proceedings of the International Conference on Intelligent Virtual Agents* (pp. 256–268).

Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science, 323,* 951–954.

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

* Saerbeck, M., Schut, T., Bartneck, C., & Janse, M. D. (2010). Expressive robots in education: Varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the 28th International Conference on Human factors in Computing Systems* (pp. 1613–1622).

Salaberry, M. R. (2001). The use of technology for second language learning and teaching: A retrospective. *The Modern Language Journal, 85,* 39–56.

Schmidt, R. W. (1990). The role of consciousness in second language. *Applied Linguistics*, *11*(2), 129–158.

* Schodde, T., Bergmann, K., & Kopp, S. (2017). Adaptive robot language tutoring based on Bayesian knowledge tracing and predictive decision-making. In *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 128–136).

Sciutti, A., Rea, F., & Sandini, G. (2014). When you are young, (robot's) looks matter. Developmental changes in the desired properties of a robot friend. In *2014 RO-MAN: The 23rd IEEE International Symposium on Robot and Human Interactive Communication* (pp. 567-573).

Scofield, J., Hernandez-Reif, M., & Keith, A. B. (2009). Preschool children's multimodal word learning. *Journal of Cognition and Development*, *10*(4), 306–333.

Service, E. (1992). Phonology, working memory, and foreign-language learning. *The Quarterly Journal of Experimental Psychology*, *45A*(1), 21–50.

Shahid, S., Krahmer, E., & Swerts, M. (2014). Child–robot interaction across cultures: How does playing a game with a social robot compare to playing a game alone or with a friend?. *Computers in Human Behavior*, *40*, 86-100.

Sheng, L., Lam, B. P. W., Cruz, D., & Fulton, A. (2016). A robust demonstration of the cognate facilitation effect in first-language and second-language naming. *Journal of Experimental Child Psychology*, *141*, 229-238.

* Shin, J., & Shin, D.-H. (2015). Robot as a facilitator in language conversation class. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human Robot Interaction Extended Abstracts* (pp. 11–12).

Shultz, M. (2007). Comparing test searches in PubMed and Google Scholar. *Journal of the Medical Library Association, 95,* 442–453.

Singer, I., & Gerrits, E. (2015). The effect of playing with tablet games compared with real objects on word learning by toddlers. In *Proceedings of the International Conference ICT for Language Learning* (pp. 2–5).

Smakman, M. (2018). *Moral considerations regarding robots in education: A systematic literature review.* Unpublished working paper, Hogeschool Utrecht, Utrecht, the Netherlands.

Smith, L. B. (2005). Action alters shape categories. *Cognitive Science*, *29*, 665–679.

Smith, L., & Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial Life*, *11*(1-2), 13-29.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558-1568.

Stanovich, K. E. (2009). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Journal of Education*, *189*(1-2), 23-55.

Stevens, C., & Bavelier, D. (2012). The role of selective attention on academic foundations: A cognitive neuroscience perspective. *Developmental cognitive neuroscience*, *2*, S30-S48.

Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: Evolving generalization through selective processing. *Nature Neuroscience*, *16*(2), 139–145.

Stockwell, G. (2008). Investigatig learner preparedness for and usage patterns of mobile learning. *ReCALL, 20,* 253–270.

Takacs, Z. K., Swart, E. K., & Bus, A. G. (2015). Benefits and pitfalls of multimedia and interactive features in technology-enhanced storybooks: A meta-analysis. *Review of Educational Research, 85,* 698–739.

Tanaka, F., & Ghosh, M. (2011). The implementation of care-receiving robot at an English learning school for children. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (pp. 265–266).

* Tanaka, F., & Matsuzoe, S. (2012). Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction, 1,* 78–95.

Tanaka, F., Cicourel, A., & Movellan, J. R. (2007). Socialization between toddlers and robots at an early childhood education center. In *Proceedings of the National Academy of Sciences* (pp. 17954–17958).

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

Tomasello, M., & Vaish, A. (2013). The origins of human cooperation and morality. *Annual Review of Psychology, 64*, 231-255.

Topping, K., Hill, S., McKaig, A., Rogers, C., Rushi, N., & Young, D. (1997). Paired reciprocal peer tutoring in undergraduate economics. *Innovations in Education & Training International, 34*(2), 96–113.

Toumpaniari, K., Loyens, S., Mavilidi, M., & Paas, F. (2015). Preschool children's foreign language vocabulary learning by embodying words through physical activity and gesturing. *Educational Psychology Review*, *27*, 445–456.

* Uluer, P., Akalın, N., & Köse, H. (2015). A new robotic platform for sign language tutoring: Humanoid robots as assistive game companions for teaching sign language. *International Journal of Social Robotics, 7,* 571–585.

Unsworth, S. (2016). Quantity and quality of language input in bilingual language development. *Bilingualism across the lifespan: Factors moderating language proficiency*, 103-122.

Uttal, D. H., Scudder, K. V, & DeLoache, J. S. (1997). Manipulatives as symbols: A new perspective on the use of concrete objects to teach mathematics. *Journal of Applied Developmental Psychology*, *18*, 37–54.

* van den Berghe, R., van der Ven, S., Verhagen, J., Oudgenoeg-Paz, O., Papadopoulos, F., & Leseman, P. (2018). Investigating the effects of a robot peer on L2 word learning. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 267–268).

van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., van der Ven, S., & Leseman, P.P.M. (2019). Social robots for language learning: A review. *Review of Educational Research, 89*(2), 259-295.

van den Berghe, R., Kramer, E., Oudgenoeg-Paz, O., Leseman, P., de Haas, M., de Wit, J., ... & Willemsen, B. (2018). *D7.6 Integrated report and recommendations.* Deliverable for the EU Horizon 2020 committee.

Verhagen, J., Boom, J., Mulder, H., De Bree, E., & Leseman, P.P.M. (in press). Reciprocal relationships between nonword repetition and vocabulary during the preschool years. *Developmental Psychology.*

Verhagen, J., de Bree, E., Mulder, H., & Leseman, P. (2017). Effects of vocabulary and phonotactic probability on 2-year-olds' nonword repetition. *Journal of Psycholinguistic Research*, *46*(3), 507–524.

Verhagen, J., Grassmann, S., & Küntay, A.K. (2017). Monolingual and bilingual children's resolution of referential conflicts; Effects of bilingualism and relative language proficiency. *Cognitive Development, 41,* 10-18.

Verhagen, J.*, van den Berghe, R.*, Oudgenoeg-Paz, O., Küntay, A., & Leseman, P. (in press). Children's reliance on the non-verbal cues of a robot versus a human. *PLOS ONE.*

Vertovec, S. (2007). Super-diversity and its implications. *Ethnic and Racial Studies*, *30*(6), 1024-1054.

Vogt, P., De Haas, M., De Jong, C., Baxter, P., & Krahmer, E. (2017). Child-robot interactions for second language tutoring to preschool children. *Frontiers in Human Neuroscience*, *11*, 73.

Vogt, P., van den Berghe, R., de Haas, M., Hoffmann, L., Kanero, J., Mamus, E., ... Kumar Pandey, A. (2019). Second language tutoring using social robots: A large-scale study. In *Proceedings of the 2019 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 497-505).

Vygotsky, L. (1978). Interaction between learning and development. In M. Cole, V. John-Steiner, S. Scriber, & E. Souberman (Eds.), *Mind in society: The development of higher psychological processes* (pp. 79–97). Cambridge: Harvard University Press.

Wainer, J., Feil-Seifer, D. J., Shell, D. A., & Matari, M. J. (2007). Embodiment and human-robot interaction: A task-based perspective. In *Proceedings of the International Conference on Human-Robot Interaction* (pp. 872–877).

Wallbridge, C. D., Lemaignan, S., & Belpaeme, T. (2017). Qualitative review of object recognition techniques for tabletop manipulation. In *Proceedings of the 5th International Conference on Human Agent Interaction* (pp. 359-363). ACM.

* Wang, Y. H., Young, S. S.-C., & Jang, J.-S. R. (2013). Using tangible companions for enhancing learning English conversation. *Journal of Educational Technology & Society, 16,* 296–309.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of personality and social psychology*, *99*(3), 410.

Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Educational Research*, *13*(1), 21-39.

Wellsby, M., & Pexman, P. M. (2014). Developing embodied cognition: Insights from children's concepts and language processing. *Frontiers in Psychology, 5.*

Wolter, B. (2006). Lexical network structures and L2 vocabulary acquisition: The role of L1 lexical/conceptual knowledge. *Applied Linguistics*, *27*(4), 741–747.

Yarrow, F., & Topping, K. J. (2001). Collaborative writing: The effects of metacognitive prompting and structured peer interaction. *British Journal of Educational Psychology*, *71*(2), 261-282.

* You, Z.-J., Shen, C.-Y., Chang, C.-W., Liu, B.-J., & Chen, G.-D. (2006). A robot as a teaching assistant in an English class. In *Proceedings of the Sixth International Conference on Advanced Learning Technologies* (pp. 87–91).

Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., ... Yukhymenko, M. (2012). Our princess is in another castle: A review of trends in serious gaming for education. *Review of Educational Research, 82,* 61–89.

Nederlandse samenvatting

Nieuwe digitale technologieën spelen een steeds grotere rol in het dagelijkse leven en worden ook steeds vaker in het onderwijs ingezet. Een van de nieuwste ontwikkelingen op dit gebied is het gebruik van sociale robots voor instructiedoeleinden in het onderwijs, ter ondersteuning of zelfs vervanging van de leerkracht. In dit proefschrift wordt onderzoek gerapporteerd naar het gebruik van sociale robots om jonge kinderen een tweede taal (T2) te leren. Voordelen van sociale robots boven andere vormen van digitale technologie, zoals tablets, zijn dat zulke robots vaak een menselijk uiterlijk hebben wat interactie zou kunnen stimuleren, non-verbale communicatievormen zoals ondersteunende gebaren kunnen gebruiken en de fysieke wereld met hun menselijke interactiepartners delen. In het onderwijs zou dit tot meer actieve en interactieve leersituaties kunnen leiden en tot grotere betrokkenheid van leerlingen bij het leerproces. Vooral deze laatste aspecten zijn van belang, omdat jonge kinderen de beginselen van taal leren door te handelen in en taalinput te krijgen over hun fysieke omgeving. Dankzij deze voordelen zouden robots kunnen worden ingezet in het T2-onderwijs.

Er is een groeiende behoefte aan effectief T2-onderwijs voor jonge kinderen om twee redenen. Enerzijds is er door de globaliserende samenleving een toenemende behoefte aan het op jonge leeftijd aanbieden van T2-onderwijs, met name Engelstalig onderwijs. Anderzijds is er steeds vaker sprake van zogeheten 'hyperdiversiteit' op scholen, waarbij veel kinderen een andere moedertaal hebben dan de taal die op school gebruikt wordt. Voor de ontwikkeling en het welzijn van deze kinderen is het belangrijk dat zij ondersteund worden bij het leren van zowel de schooltaal als hun moedertaal.

Ondanks de beloften en ook het enthousiasme bij scholen om met nieuwe technologieën aan de slag te gaan, is er nog maar weinig bekend over de effectiviteit van sociale robots in (tweede)taalonderwijs. De weinige studies die tot nu toe zijn uitgevoerd naar T2-woordleren vinden geen eenduidige, en soms zelfs contra-intuïtieve resultaten. Zo is er in onderzoek een effect gevonden op het woordleren van één enkele, kortdurende sessie met de robot, terwijl een reeks van sessies met de robot juist geen leereffect liet zien. Normaal gesproken zijn woordenschatinterventies juist effectiever naarmate ze uit meer sessies bestaan. Er is weinig tot geen onderzoek geweest naar diverse factoren die de effectiviteit van de robot kunnen beïnvloeden, zoals de nieuwigheid van de robot, het aantal sessies waarin kinderen met de robot leren, en individuele verschillen tussen kinderen in bijvoorbeeld het algemeen leervermogen of afleidbaarheid. Ook is nog niet duidelijk of de voordelen die robots in theorie hebben ten opzichte van andere vormen van technologie, zoals tablets, ook daadwerkelijk helpen bij het leren.

Het onderzoek dat beschreven is in dit proefschrift, is uitgevoerd binnen het L2TOR-project. Het doel van dit project was om een robot te ontwikkelen die jonge kinderen kan ondersteunen bij het leren van woorden in een T2. Binnen het L2TOR-project werd een lessenserie ontwikkeld die aansluit bij de huidige kennis over woordenschatinterventies bij jonge kinderen. In deze lessenserie speelden kleuters samen met een NAO-robot taalspelletjes op een tablet. In dit proefschrift wordt een aantal studies beschreven die gebruik maakten van dit systeem en waarin Nederlandse kleuters Engelse woorden leerden. De doelen van dit proefschrift waren om (1) de verschillende resultaten uit eerder uitgevoerd onderzoek naar robot-ondersteunde taalinstructie samen te brengen en te zien waar de open vragen liggen; (2) het gebruik van tablets ten opzichte van fysieke objecten binnen T2-onderwijs met robots te onderzoeken; (3) de toegevoegde waarde van robots voor T2-woordleren te onderzoeken; en (4) te onderzoeken of individuele verschillen tussen kinderen in hun taalvaardigheid, aandacht en perceptie van de robot een rol spelen in de mate waarin de robot jonge Nederlandse kinderen helpt bij het leren van Engelse woorden.

Hoofdstuk 2 bevat een overzichtsstudie van de recente onderzoeksliteratuur over robot-ondersteund (taal)onderwijs. Het doel van de studie was om zowel de resultaten uit eerder uitgevoerd onderzoek als open vragen in kaart te brengen. De tot nu toe uitgevoerde onderzoeken met voldoende methodologische zeggingskracht, 33 in totaal, verschillen sterk van elkaar in onder meer de groep proefpersonen die is onderzocht, het type robot dat is gebruikt, de taal die werd onderwezen, en de onderwijsmethode. Allereerst kwam uit deze review naar voren dat er sterk wisselende resultaten worden gevonden met betrekking tot leeropbrengsten, met name bij het leren van woorden in een T2. Eenduidiger was het positieve effect van het gebruik van robots op de motivatie van leerlingen. Hierbij moet de kanttekening geplaatst worden dat dit effect alleen gevonden leek te worden zolang de robot nog nieuw was voor de leerlingen, en dat dit effect geleidelijk verdween wanneer leerlingen bekender raakten met de robot. Tot slot bleek het sociale gedrag van de robot een complexe rol te spelen. Sociaal gedrag van de robot bleek in sommige onderzoeken te kunnen bijdragen aan het leren, maar in andere onderzoeken juist afleidend of intimiderend te werken. Eerder onderzoek laat dus zien dat het lastig is een goede balans te vinden tussen sociaal of motiverend gedrag van een robot enerzijds, en afleidend gedrag anderzijds. Bovendien leek de optimale vorm van sociaal gedrag te verschillen tussen leerlingen. Samengevat bleek dat veel van de voordelen die robots in theorie hebben, in praktijk nog niet makkelijk kunnen worden gerealiseerd. Er is nog veel onderzoek nodig naar het optimale ontwerp van T2-lessen met robots en welke rol robots in deze lessen precies moeten vervullen.

In Hoofdstuk 3 werd één zo’n ontwerpaspect onderzocht, namelijk het gebruik van tablets. Tablets worden vaak gebruikt in T2-lessen met robots wegens beperkingen van de robot op het gebied van spraak- en objectherkenning, maar het is onduidelijk hoe dit het leren beïnvloedt. Vanuit de theorie van *embodied cognition* wordt verwacht dat kinderen minder leren wanneer ze virtuele voorwerpen manipuleren op tablets dan wanneer ze fysieke objecten manipuleren, omdat de sensomotorische ervaring die ze opdoen wezenlijk anders is. Zo zouden kinderen bij het leren van het woord 'heavy' baat kunnen hebben bij het vasthouden van een relatief zwaar object ten opzichte van het verslepen van een virtueel (en dus gewichtsloos) object. In het experiment dat beschreven wordt in Hoofdstuk 3, werd het effect van het gebruik van 3D-beelden van speelgoeddieren op tablets voor T2-woordleren vergeleken met fysieke speelgoeddieren die om bijvoorbeeld het aspect *zwaar* goed waarneembaar te maken extra waren verzwaard. Er werden echter geen verschillen gevonden: kinderen die een les volgden waarin zij Engelse woorden leerden aan de hand van het verslepen en aanklikken op een tablet leerden niet meer of minder Engelse woorden dan kinderen die een les volgden waarin zij fysieke objecten moesten optillen en verplaatsen. Dit zou kunnen komen doordat kinderen al bekend waren met de concepten (zoals 'zwaar')in hun moedertaal, en daardoor niet het concept hoefden te ervaren (door iets zwaars op te tillen) om het woord te leren. Voor het ontwerpen van T2-lessen met robots waarbij vanwege de technische beperkingen van de robot een tablet gebruikt moet worden, betekent dit resultaat dat werken met de tablet niet per se nadelig is voor het leren.

In de Hoofdstukken 4 en 5 werd de toegevoegde waarde van de robot als leermaatje onderzocht. Samen leren met iemand anders kan namelijk helpen bij het leren. In het experiment dat beschreven wordt in Hoofdstuk 4 werden drie condities vergeleken: (1) kinderen die zonder maatje Engelse woorden leerden; (2) kinderen die samen met een klasgenootje leerden; en (3) kinderen die samen met een robot leerden. Een volwassene vertelde een verhaal waarin de Engelse woorden waren verwerkt en gaf de kinderen (en de robot) instructies om opdrachten uit te voeren om met de woorden te oefenen. Er werden geen verschillen in leereffecten gevonden tussen de drie condities op een woordkennistest die direct na de les werd afgenomen. Bij een test die een week later werd afgenomen, presteerden kinderen die alleen, dus zonder een klasgenootje of robot, hadden geleerd zelfs beter dan de kinderen in de twee andere condities. Ook wat betreft de beleving van de les werden er geen verschillen gevonden tussen de drie condities: de kinderen vonden de les even leuk. Er moet hierbij opgemerkt worden dat de les vrij gestructureerd was en dat de volwassene een belangrijke rol speelde. Wellicht profiteren kinderen wel van de robot als leermaatje als ze meer zijn aangewezen op de robot en meer kunnen profiteren van de kennis die de robot biedt.

De studie die in Hoofdstuk 5 besproken werd met betrekking tot de toegevoegde waarde van robots, was de hoofdstudie van het L2TOR-project. Bijna 200 kinderen volgden individueel een goed opgebouwde reeks van zeven lessen waarin zij leeftijdsadequate Engelse woorden leerden op het gebied van ruimtelijke en rekenkundige kennis. Het gebruik van een lessenserie maakte het mogelijk het effect van de robot over meerdere sessies te onderzoeken. Het was een van de eerste onderzoeken in het veld van mens-robotinteractie dat op deze schaal en met een gerandomiseerde onderzoeksopzet de effectiviteit van robots voor onderwijsdoeleinden heeft onderzocht. Er waren drie experimentele condities, waarin kinderen woorden leerden met (1) alleen een tablet, (2) een tablet en de robot, of (3) een tablet en de robot die iconische gebaren maakte voor ieder T2 woord. Bovendien was er een controleconditie waarin kinderen geen T2-woorden leerden, maar een niet-gerelateerde taak uitvoerden met de robot (namelijk een dans). De verwachting dat de robot en met name de iconische gebaren van de robot toegevoegde waarde zouden hebben voor het leren van Engelse woorden, werd niet bevestigd. Er werden geen verschillen gevonden tussen de drie experimentele condities. Kinderen leerden dus niet meer woorden in het Engels wanneer zij de les kregen aangeboden met de robot en een tablet dan wanneer zij de lessen zonder robot op een tablet volgden. Ook wanneer de robot iconische gebaren gebruikte leidde dit niet tot hogere leeropbrengsten. Kinderen leerden in de experimentele condities wel meer dan in de controleconditie, wat laat zien dat ze leerden van de lessen. Het gebrek aan verschillen tussen de drie experimentele condities zou kunnen komen doordat in alle condities de kinderen sterk gericht waren op de tablet, waarop de meeste handelingen uitgevoerd moesten worden. Dit was helaas onvermijdelijk, vanwege de beperkingen op het gebied van spraak- en objectherkenning. De resultaten van Hoofdstukken 4 en 5 sluiten daarmee aan bij de conclusie van de overzichtsstudie in Hoofdstuk 2, dat de voordelen van robots voor T2-onderwijs vooral nog in theorie bestaan en nog niet in de praktijk.

In de Hoofdstukken 5 en 6 werd aandacht besteed aan individuele verschillen tussen kinderen die de effectiviteit van de robot als taaltutor zouden kunnen beïnvloeden. In Hoofdstuk 5 werd onderzocht of verschillen in woordenschat in de moedertaal (het Nederlands), fonologisch geheugen en selectieve aandacht als zogenaamde moderatorvariabelen van invloed waren op de effectiviteit van de robot. Verschillen in taalvaardigheid en aandacht bleken inderdaad het effect van de robot op het woordleren te modereren. Kinderen met een grote Nederlandse woordenschat leerden meer Engelse woorden wanneer zij met robot leerden dan wanneer zij zonder robot leerden, terwijl kinderen met een goed fonologisch geheugen juist meer leerden in de conditie zonder robot ten opzichte van de robot-ondersteunde condities. Kinderen met een goed ontwikkelde selectieve aandacht leerden meer Engelse woorden in de conditie waarin de robot iconische gebaren gebruikte, terwijl kinderen met minder goed ontwikkelde aandacht meer baat hadden bij een robot zonder gebaren. Een mogelijke verklaring is dat kinderen alleen kunnen profiteren van de extra informatie die via gebaren wordt aangeboden, als zij goede vaardigheden hebben wat betreft het richten van hun aandacht.

In Hoofdstuk 6 werd onderzocht of verschillen tussen kinderen in hoe zij denken over de robot van invloed zou kunnen zijn op de effectiviteit van de robot als taaltutor. Zowel voor als na de lessenserie werd de kinderen gevraagd welke menselijke kenmerken wel of niet van toepassing zijn op de robot, zoals eten nodig hebben, kunnen begrijpen wat iemand zegt, of blij kunnen zijn. Over het algemeen hadden de kinderen de neiging om menselijke eigenschappen toe te schrijven aan de robot, maar er waren grote individuele verschillen in de mate waarin ze dat deden. Veel kinderen waren voorafgaand aan de lessenserie geneigd om biologische eigenschappen toe te schrijven aan de robot, maar zagen de robot na de lessenserie meer als een mechanisch apparaat dat echter wel over allerlei cognitieve vaardigheden beschikte en in staat was tot begrip. De mate waarin kinderen meer menselijke eigenschappen aan de robot toeschreven na afloop van de lessenserie ten opzichte van voorafgaand aan de lessenserie bleek zwak, maar significant positief gerelateerd aan hoeveel Engelse woorden uit de lessenserie zij nog wisten enkele weken na afloop van de lessen. De resultaten die in Hoofdstukken 5 en 6 worden gerapporteerd laten zien dat kinderen verschillend reageren op de robot en dat deze verschillen gerelateerd zijn aan de leerwinst die zij boeken in de robot-ondersteunde T2-lessen.

Uit het onderzoek van dit proefschrift blijkt dat er nog veel technologische beperkingen zijn die eerst moeten worden opgelost, voordat robots effectief ingezet kunnen worden als leraar of leermaatje in T2-onderwijs. Zelfs als deze technologische beperkingen kunnen worden opgelost, is het de vraag of het ontwikkelen van sociale robots die de rol van leraar of leermaatje kunnen overnemen wel de beste aanpak is voor dit veld. Volwaardige deelname aan (leer)interacties met mensen, kinderen in het bijzonder, vraagt meer van de robot – met name begrip en empathie – en het is twijfelachtig of deze kenmerkende aspecten van menselijke interactie op afzienbare termijn ingebouwd kunnen worden in sociale robots. Zolang dit niet zo is, zal de robot snel door de mand vallen. Werken aan verbetering van oppervlakkige interactiekwaliteiten van sociale robots, zonder begrip en empathie, kan mensen, en met name kinderen, misleiden en roept ethische vragen op. Wellicht kan de onderzoeks- en ontwikkelingsagenda zich beter richten op kwaliteiten die robots nu al hebben en waarmee ze leraren kunnen aanvullen. Dit kan bijvoorbeeld ondersteuning van het leerproces betreffen bij taken met veel herhaling of waarvoor weinig interactie en begrip nodig is. In tegenstelling tot leraren hebben robots eindeloos geduld en eindeloos de tijd om bepaalde oefeningen met kinderen te herhalen. Daarnaast is het belangrijk om kinderen meer te leren over de aard van robots en hoe ze werken. Dit wordt steeds belangrijker omdat er steeds meer robots in werk, zorg en privéleven zullen komen.

Samengevat blijkt dat robots weliswaar potentieel hebben als nieuwe digitale technologie in het onderwijs, maar dat ze dit op dit moment nog niet waar kunnen maken. Dit proefschrift leert dat: (1) het gebruik van tablets niet per se nadelig is voor T2-woordleren, maar dat robotlessen waarin de aandacht van leerlingen teveel gericht is op de tablet kunnen resulteren in een gebrek aan voordelen van robots; (2) verbaal en non-verbaal gedrag van robots met zorg ontwikkeld moeten worden om kinderen van de robot te laten profiteren zonder ze erdoor te laten afleiden; en (3) er rekening gehouden moet worden met individuele verschillen in vaardigheden en perceptie bij het ontwikkelen van robot-ondersteund onderwijs. De belangrijkste beperkingen van het onderzoek dat in dit proefschrift is gerapporteerd, betreffen het robot-tablet systeem: niet alle voordelen die robots in theorie boven andere vormen van technologie zouden kunnen hebben, konden ook daadwerkelijk worden benut. Dit zou mede kunnen verklaren dat er geen toegevoegde waarde van het gebruik van de robot werd gevonden. Verwant hieraan is dat er geen vergelijking kon worden gemaakt tussen een robot zonder tablet en andere technologieën. Niet duidelijk is hoe effectief een robot zonder tablet is ten opzichte van andere vormen van technologie zoals tablets. In de toekomst moeten robots verder ontwikkeld worden om alle mogelijkheden eruit te halen die ze nu vooral nog in theorie hebben. Daarnaast moet ook vooral gekeken worden naar hoe robots verschillen van leraren en hoe ze leraren kunnen aanvullen. De onderzoeken die in dit proefschrift zijn beschreven, vormen de eerste stappen om effectievere robots voor T2-onderwijs te ontwikkelen en het potentieel van robots waar te maken.

Dankwoord (acknowledgments)

Er zijn zoveel mensen die ik wil bedanken! Mijn promotietraject is een hele bijzondere tijd geweest waarvan ik volop heb kunnen genieten, dankzij jullie.

Paul, Josje en Ora, wat ben ik blij dat ik jullie als begeleiders had. Regelmatig is me gevraagd of ik niet gillend gek werd met een promotietraject van 2,5 jaar, en jullie waren de reden dat ik steevast zei "het is goed te doen als je fijne mensen om je heen hebt". Mijn promotietijd was een tijd van persoonlijke hoogte- en dieptepunten. Bedankt voor jullie steun tijdens de mooie en vooral ook de moeilijke momenten. Ik heb bijzonder veel van jullie geleerd. Paul, regelmatig zat ik als net afgestudeerde taalwetenschapper met grote ogen te kijken als je weer aankwam met een theorie die ik niet kende. Bedankt voor het vertrouwen dat je uitstraalde, wat ik zeker in het begin goed kon gebruiken, en voor alle interessante gesprekken over het gebruik van robots in het onderwijs. Josje, het beginnen van een PhD voelde als een sprong in het diepe, maar jij was een soort baken waar ik me aan kon vasthouden. Je hebt me altijd op een fijne manier uitgedaagd bij het uitdenken en opschrijven van onderzoek. Ik ga onze gezamenlijke treinritjes missen, al zullen we vast nog regelmatig afspreken voor een kopje thee op een willekeurige plek. Ora, dingen leken altijd op de een of andere manier minder lastig als ik even met jou had gepraat. Was het chaos in mijn hoofd of had ik een probleem waar ik zelf niet uit kwam, hoefde ik maar even naar jou toe te gaan en dan kwamen we wel tot een oplossing. Ik ben je ontzettend dankbaar voor hoeveel je in het afgelopen jaar in het bijzonder hebt gedaan om te zorgen dat ik genoeg tijd had om aan mijn proefschrift te werken.

I would also like to thank all people involved in the L2TOR project. We are with too many to list everybody by name, but please know that I am very grateful for all your input and that I have very much enjoyed our meetings, in real-life and through Skype. We have had a very fruitful collaboration over the last couple of years, and I hope to work with you again in the future. In particular I would like to thank Mirjam, Jan, Bram, and Thorsten. With colleagues like you, you don’t need that much more to be happy (except of course for coffee, stroopwafels, and turtles).

Sanne, jij was ook een belangrijk onderdeel van het Utrechtse L2TOR team. Jouw kritische blik was ontzettend waardevol bij het opzetten van de onderzoeken. Erica, bedankt voor je hulp bij de technische kant van het project. Onze maandagmiddagen waarin we met de robot leerden te werken, hebben me enorm geholpen om wat vertrouwen te krijgen dat ik met robots kan werken!

Ook dank aan de vele anderen op de universiteit. Mijn lieve kamergenootjes van de taalkamer: jullie hebben laten zien dat gezelligheid en productiviteit zeker samen kunnen gaan. Ook dank aan de vele anderen van de afdeling Orthopedagogiek, voor de gezellige lunches en praatjes tussen alle werkzaamheden door. In het bijzonder wil ik degenen van de Language Meetings bedanken voor de fijne bijeenkomsten. Daarnaast wil ik natuurlijk mijn schrijfgroepje nog even noemen. Regelmatig liep ik met een hoofd vol chaos naar the Village Uithof om na onze bijeenkomst weer vol nieuwe plannen, goede moed en een lijst met doelen (en niet-doelen!) naar kantoor terug te keren.

Tijdens het L2TOR project zijn er vele studenten en assistenten geweest, die hun stage of scriptie bij ons deden of kortere of langere tijd voor ons hebben gewerkt. Bedankt Annelies, Bente, David, Deborah, Ellis, Esmee V., Hanneke, Hugo, Ilse, Laurette, Lisa, Loes, Lotte, Michelle Ze., Michelle Zo., Madée, Peggy, Robin, Sirkka, Sam en Veerle.

Ook de studenten en assistenten die vanuit Tilburg University met de hoofdstudie hebben meegeholpen: Annabella, Chani, Laura, Marije, Pieter, Reinjet en Sabine. In het bijzonder wil ik Esmee K. en Bram bedanken. Jullie hebben in de laatste maanden van L2TOR niet alleen heel veel voor me gedaan waardoor ik tijd had om te schrijven, jullie hebben het ook tot de gezelligste maanden in het Langeveld gemaakt!

Mijn proefschrift had niet tot stand kunnen komen zonder de vele kinderen, ouders en scholen die hebben meegewerkt aan het onderzoek. Tijdens de dataverzameling had ik regelmatig een grote lach op mijn gezicht vanwege de mooie uitspraken die kinderen deden richting en over Robin de robot (waarbij mijn favoriet toch wel is "ja, Robin de robot groeit als hij de korstjes van zijn brood opeet"). Mijn dank aan jullie deelname en medewerking is groot!

Verder wil ik mijn lieve vrienden en familie bedanken. Ik denk dat jullie het grootste gedeelte van de tijd geen idee hadden waar ik nou precies mee bezig was ("iets met robots?"), maar dat maakte jullie vertrouwen in mij er niet minder om. Bedankt voor jullie steun! De vele gezellige etentjes, spelletjesavonden en koffietjes zorgden ervoor dat het geen enkel probleem was om in de avonden en weekenden mijn aandacht op iets anders te vestigen dan mijn promotieonderzoek. Giulia en Mirjam, bedankt dat jullie mij bijstaan als paranimfen. Papa en mama, naast dat jullie het voor mij mogelijk hebben gemaakt om te gaan studeren, hebben jullie me gestimuleerd om te gaan doen wat ik leuk vind. Ik ben ontzettend blij dat ik daardoor op een plek ben terecht gekomen waar ik zo van geniet. Leonie, dankjewel voor het helpen ontwerpen van de omslag van mijn boekje.

En dan tot slot: lieve Frank. Zonder jou durfde ik al niet eens aan de onderzoeksmaster te beginnen, laat staan dat ik een PhD was gaan doen. Dankjewel voor je vertrouwen, je rust op de momenten waarop ik weer eens niet wist wat ik moest doen, en voor alles wat je hebt gedaan om ervoor te zorgen dat ik mijn proefschrift af kon maken. Wat is er een hoop op ons afgekomen de afgelopen 2,5 jaar, maar het is duidelijk dat we samen alles aankunnen. Terugkijkend herbeleef ik alle mooie momenten en wat ben ik trots dat mijn nieuwe naam op de voorkant van dit boekje prijkt! Ons volgende avontuur staat al voor de deur en ik kan niet wachten om dat samen met jou te gaan beleven!

P.s. En natuurlijk Harry, bedankt voor de motivatie om mijn proefschrift af te ronden.

About the author

Rianne van den Berghe was born on 20 March 1993 in Schagen, the Netherlands. She obtained her high school degree (gymnasium) in 2011 from the Regius College in Schagen. She went on to study Liberal Arts and Sciences at Utrecht University, with a major in Linguistics. She discovered her interest in research while working at the Babylabs of Utrecht University and Radboud University Nijmegen. In 2016, she obtained her master degree cum laude from Utrecht University, having completed the research master 'Linguistics: the study of the language faculty'. During her studies, she focused on language acquisition, discourse processes, and theory of mind, and did an internship at Turku University, in Finland, on effects of bilingualism in word recognition processes.

In June 2016, Rianne started her PhD at the department of Special Education at Utrecht University. She worked at the L2TOR project, which was financed by the European Union Horizon2020 program, awarded to a consortium led by Tony Belpaeme (Ghent University) and Paul Vogt (Tilburg University). During her PhD, she organized departmental language meetings, gave guest lectures, and supervised master theses. She also presented her research at national and international conferences and took up the role of treasurer for the *Werkverband Amsterdamse Psycholinguïsten*. After her PhD, she started as a postdoctoral researcher at Utrecht University, investigating the use of robots to aid children with autism spectrum disorder in learning technical and social skills, and as a junior researcher at Radboud University Nijmegen, investigating simulation processes during auditory processing of literary narratives.

Publications

Belpaeme, T., Vogt, P., **van den Berghe, R.**, Bergmann, K., Göksun, T., De Haas, M., ... & Papadopoulos, F. (2018). Guidelines for designing social robots as second language tutors. *International Journal of Social Robotics*, 1-17.

de Wit, J., Pijpers, L., **van den Berghe, R.**, Krahmer, E., & Vogt, P. (2019). Why UX research matters for HRI: The case of tablets as mediators. In *Workshop on the Challenges of Working on Social Robots that Collaborate with People at CHI 2019.*

Rintjema, E., **van den Berghe, R.**, Kessels, A., de Wit, J. & Vogt, P. (2018). A robot teaching young children a second language: The effect of multiple interactions on engagement and performance. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 219-220).

van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., van der Ven, S., & Leseman, P.P.M. (2019). Social robots for language learning: A review. *Review of Educational Research, 89*(2), 259-295.

van den Berghe, R., van der Ven, S.H.G., Verhagen, J., Oudgenoeg-Paz, O., Fotios, Papadopoulos & Leseman, P.P.M. (2018). Investigating the effects of a robot peer on L2 word learning. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 267-268).

Verhagen, J.*, **van den Berghe, R.***, Oudgenoeg-Paz, O., Küntay, A., & Leseman, P. (in press). Children's reliance on the non-verbal cues of a robot versus a human. *PLOS ONE.*

Vogt, P., **van den Berghe, R.**, de Haas, M., Hoffman, L., Kanero, J., Mamus, E., … & Pandey, A.K. (2019). Second language tutoring using social robots: A large-scale study. In *Proceedings of the 2019 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 497-505).

Vlaar, R., Verhagen, J., Oudgenoeg-Paz, O., & Leseman, P. P. M. (2017). Comparing L2 word learning through a tablet or real objects: What benefits learning most?. In *Proceedings of the Workshop R4L, at ACM/IEEE HRI 2017.*

Wallbridge, C. D., **van den Berghe, R.**, Hernández Garcia, D., Kanero, J., Lemaignan, S., Edmunds, C., & Belpaeme, T. (2018). Using a robot peer to encourage the production of spatial concepts in a second language. In *Proceedings of the 6th International Conference on Human-Agent Interaction* (pp. 54-60).