# I LIKE THE WAY YOU MOVE

Robots that gesture, and their potential as second language tutors for children

J.M.S. de Wit

# I like the way you move

## Robots that gesture, and their potential as second language tutors for children

Jan de Wit

Robots that gesture, and their potential as second language tutors for children

Jan Mark Sander de Wit
PhD Thesis
Tilburg University, 2022

TiCC PhD Series No. 78
ISBN 978-94-6458-015-0

Design: Renée de Leau | www.reneedeleau.com
Printing: Ridderprint | www.ridderprint.nl

# Robots that gesture, and their potential as second language tutors for children

PROEFSCHRIFT

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University
op gezag van de rector magnificus, prof. dr. W.B.H.J. van de Donk,
in het openbaar te verdedigen ten overstaan van
een door het college voor promoties aangewezen commissie
in de Aula van de Universiteit op vrijdag 28 januari 2022 om 10.00 uur

door
**Jan Mark Sander de Wit,**
geboren te Drunen.

| | |
|---|---|
| Promotor: | prof. dr. E.J. Krahmer (Tilburg University) |
| Copromotor: | dr. P.A. Vogt (Hanzehogeschool Groningen) |
| Leden promotiecommissie: | prof. dr. T. Belpaeme (Ghent University)<br>dr. M.M.A. de Graaf (Utrecht University)<br>prof. dr. S.D. Kelly (Colgate University)<br>prof. dr. P. Markopoulos (Eindhoven University of Technology)<br>prof. dr. K.J. Rohlfing (Paderborn University)<br>prof. dr. M.G.J. Swerts (Tilburg University) |

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

For mom, my best friend, and the strongest, kindest person I know.

# Contents

1

Chapter 1

# Introduction

"I spy with my little eye... *bird*."

The girl gets up, stretches her arms, and starts moving them up and down gracefully to depict the act of flying. I am in the middle of conducting our first study (Chapter 3), and this was the first time I had seen one of the children get up and mimic one of the robot's gestures. That is when I realized that there is something magical about how these gestures can contribute to our communication with each other and, as it turns out, with robots as well.

This first study was part of the L2TOR ('el tutor') project, a European Commission-funded collaboration between several universities and industry partners across Europe (Belpaeme et al., 2015). The aim of this project was to investigate the potential role of *social robots* — robots that look human-like, and are designed to engage with people in everyday interactions (Bartneck & Forlizzi, 2004; Duffy, 2003) — as tutors for second language (L2) learning. There is a need for innovative technological tools in education, because the average number of students per teacher is increasing. This limits the teachers' ability to accommodate their students' individual needs, for example regarding their preferred learning style (Blatchford & Russell, 2020). The L2TOR project sought to answer the question whether social robots can fill this need for technological support in education and, if so, how we can design interactions in such a way that social robots can be effective as second language tutors.

Compared to alternative technologies that can help provide individual tutoring (e.g., tablets), social robots are physically present in the context where learning takes place. This could enable them to provide additional social support and communicate with the student in a natural, human-like way, which has been argued to be conducive to learning (Saerbeck et al., 2010). These communicative skills, as well as abilities such as detecting the learner's emotional state, are together referred to as a robot's perceived *social intelligence* (Fong et al., 2003). An important component of this social intelligence is the ability to use non-verbal communication, including the use of manual (hand) gestures. Therefore, in the current thesis we focus on studying the effects of a social robot's use of manual gestures to facilitate the learning process.

To support our communication with others, we tend to spontaneously make use of different types of hand gestures (Hostetter, 2011; McNeill, 1992; Rohlfing et al., 2012). In this thesis, we will focus mostly on *deictic* gestures, also known as pointing gestures, *iconic* gestures, and *metaphoric* gestures. Deictic gestures are used to refer to objects, people, or locations, and can be used to direct someone's visual attention (Kita, 2003). Iconic and metaphoric gestures convey meaning by eliciting a mental image that corresponds to the relevant concept (McNeill, 1985). While iconic gestures refer to concrete objects or actions (e.g., depicting a *telephone* by holding a hand up to one's ear, with the thumb and pinky finger extended), metaphoric gestures refer to abstract concepts (e.g., depicting that something is *large* by extending one's arms and spreading them far apart).

Gestures have been argued to play an important role in education as well (see, e.g., Roth, 2001, for a review). They have been shown to aid students' understanding of lesson content, and can help retain newly acquired knowledge over time (Alibali & Nathan, 2007; Cook et al., 2008). Furthermore, children's attention appears to be drawn more to the object of learning if a teacher uses gestures, which can lead to better learning outcomes compared to when a teacher only uses speech (Wakefield et al., 2018). Specifically in (second) language learning, iconic and metaphoric gestures can aid the learning process (Rohlfing, 2019), by 'grounding' new and unknown words in the student's existing non-linguistic knowledge or experiences (Barsalou, 2008; Hald et al., 2016), which can lead to improved second language vocabulary acquisition for adults (Kelly et al., 2009; Macedonia et al., 2011) as well as children (de Nooijer et al., 2013; Rowe et al., 2013; Tellier, 2008).

Because social robots typically have arms and hands in one form or another, they may be able to use gestures in a similar manner as human speakers do. It is therefore conceivable that the positive effects of gestures that are observed in education could apply to tutoring provided by social robots as well. Research into robot-performed gestures in general has indeed shown promising results (e.g., van Dijk et al., 2013; Yadollahi et al., 2018), although to the best of our knowledge their contribution specifically to second language tutoring has not been explored prior to the start of the L2TOR project. Next to supporting the robot's tutoring efforts and being part of a robot's socially intelligent behavior, gestures might also contribute to other aspects of the robot's social intelligence, such as being able to express emotions (e.g., J. Xu et al., 2014) and building rapport (e.g, Stolzenwald & Bremner, 2017). Displaying more complex and diverse forms of social intelligence, in turn, can result in sustained engagement during the educational interactions (Leite et al., 2013), and could make people more likely to accept social robots in their daily lives (de Graaf & Ben Allouch, 2013): two important elements for establishing a lasting impact on the field of education.

One key difference between the gesturing behavior of robots and humans, is that we tend to produce our gestures spontaneously (Hostetter & Alibali, 2008), while for robots the entire gesture production process has to be deliberately designed and implemented. This includes designing the gestures themselves — what should a gesture for a *car* look like? — as well as integrating these gestures with other modalities, such as speech, and adjusting to various contextual factors (e.g., adding variation to make the robot seem less repetitive). At the same time, the robots that are currently available are limited in their motor degrees of freedom, which means that it is impossible to have a robot copy the exact physical behavior of a human. The design decisions that have to be made when imbuing a robot with the ability to use gestures will likely have consequences, in terms of the gestures' contribution to second language learning or the degree of engagement with the robot. If, for example, the robot were to gesture too frequently, or the meaning of the gestures is unclear, they might not help children learn words in a second language, and may even have a detrimental (i.e., distracting) effect. It is therefore important to explore the design space of robot-performed gestures, in order to optimize the use of this modality, and by extension to optimally make use of the robot's physical presence. As a result, the main objective of this thesis is to study the effects of robot-performed gestures in the context of second language tutoring with children, and to explore how design decisions regarding the robot's gesture production process may influence these effects.

## 1.1 Social robots as (second language) tutors

Several literature reviews have explored the use of social robots as tools for education (Belpaeme et al., 2018; Mubin et al., 2013; Toh et al., 2016). Similarly to other technologies that have previously been introduced to the educational field, such as tablet devices and virtual agents, robots could potentially complement teachers — but certainly not replace them — as tutoring devices. A robot can tirelessly repeat content to practice as long as needed (Chang et al., 2010), and the tutoring interaction can be tailored to the needs of an individual student (e.g., Leyzberg et al., 2018). The added benefits of using social robots rather than alternative technologies are thought to lie in their physical presence in the real world, as well as their human-likeness and the resulting ability to provide social support to the student. However, it remains a challenging task to create robots that are able to engage in complex social interactions (Yang et al., 2018). More research is needed to determine whether social robots can live up to our expectations, how we can best make use of this new technology, and who can benefit from interacting with social robots, in which contexts.

Physical embodiment and presence in the physical world are said to be important benefits of social robots, compared to alternative technologies such as tablets. This allows them to observe, move around in, and interact with the environment in which learning takes place, which could further support their teaching efforts (e.g., Hood et al., 2015). The robot could, for example, refer to or manipulate objects in the environment. However, this avenue of research is still relatively underexplored, possibly because it is challenging for the types of social robots that are currently widely available to manipulate objects in complex, unconstrained environments (Kemp et al., 2007). The robot's physical presence in and of itself has been shown to stimulate social behavior from the student and result in greater learning outcomes (Belpaeme et al., 2018), as well as increase trust toward the robot, compared to when it is telepresent via a screen (Bainbridge et al., 2011). A literature study found that physically embodied and present robots were rated more positively compared to telepresent robots and virtual agents (Li, 2015). This was especially true if the robot used gestures, which therefore appears to be an important way for robots to take advantage of their presence in the physical world.

Because social robots often look human-like, people tend to *anthropomorphize* them: they assign human characteristics and behavior to them (Duffy, 2003). As a result, people want to — and expect to — communicate with social robots in a human-like way (Bartneck & Forlizzi, 2004), for example using natural language and non-verbal means of communication, such as gaze and hand gestures. If a robot is able to meet these expectations, people might be more likely to build social bonds with them (de Graaf, 2016). A robot's social behavior further includes aspects such as observing and exhibiting emotions, and the ability to establish social relationships (Fong et al., 2003). However, social behavior proves to be challenging to implement, as it relies on complex sensing and decision-making functionalities (i.e., picking up on social signals, and then reciprocating in an appropriate manner). Engaging in social interaction, especially in the long term, is therefore considered one of the grand challenges in robotics (Yang et al., 2018). If a robot does manage to build a social bond with people, this will likely support the robot's ability to act as a tutor. For example, research has shown that students who experience a social bond with a robot are more likely to exhibit social behavior, such as help-seeking, themselves (Howley et al., 2014), and experiencing rapport with a social robot has been shown to contribute to greater learning outcomes (Kory-Westlund & Breazeal, 2019).

Surveys in the domain of (second) language learning illustrate the potential of using social robots to support children and adults in acquiring new language skills (Kanero et al., 2018b; van den Berghe et al., 2019). An advantage of using robots is that they can switch between the student's primary (L1) and secondary (L2) languages. Students in a study also reported feeling less anxious and more motivated when a social robot accompanied a teacher during a second language vocabulary lesson (Alemi et al., 2015). Gestures in particular are mentioned as an important feature of social robots in this context, because they have previously been shown to support second language learning with human teachers (Kanero et al., 2018b). However, there is still a lack of conclusive evidence regarding the effectiveness of social robots as second language tutors, particularly in the long term. In addition, it is unclear how robot tutors compare to alternative technologies or human tutors, and to what extent their effectiveness is subject to individual differences between students.

It is possible that people's expectations, for example in terms of perceived social intelligence, exceed the capabilities of the robots that are currently available, resulting in a sense of disillusionment known as the *social robotics winter* (Henschel et al., 2020). When evaluating the design of the robot's behavior, it is therefore important to not only consider the effects of these design decisions on (short-term) learning outcomes, but to also investigate indicators of interest in, and relationship formation with social robots, such as the levels of engagement or involvement with the interaction, and how the robot is perceived by the people interacting with it (e.g., as a social agent, or rather an inanimate toy). This applies to the robot's use of hand gestures as well, which may not only result in better learning outcomes, but could also contribute to greater levels of engagement and change the students' perception of the robot.

### The L2TOR project

This thesis was carried out as part of the L2TOR ('el tutor') project (Belpaeme et al., 2015), a European Commission Horizon 2020 funded project that ran from 2016–2019, in which a number of European academic and industry partners collaborated to investigate whether social robots could successfully be used as second language tutors for children in kindergarten (4–6 years old, the first two grades of primary school in the Netherlands)[1]. The collaborating universities were Bielefeld University (Germany), Ghent University (Belgium), Koç University (Turkey), Plymouth University (United Kingdom), Tilburg University (the Netherlands), and Utrecht University (the Netherlands). The industrial collaborators were SoftBank Robotics (France) and Zora Robotics (Belgium).

To address several of the important outstanding questions in the field of social robotics for second language learning outlined in the previous paragraph, the project consisted of a number of studies that each focused on a particular design feature of the robot, such as adapting to the skill level of the student (Schodde et al., 2017, and Chapter 3 of this thesis), providing different types of feedback (de Haas et al., 2020), and gestures — the focus of this thesis. In addition, different aspects of the interaction between the children and the robot were studied, including not only the resulting learning outcomes, but also engagement with the task and the robot, the degree to which children anthropomorphized the robot (van den Berghe, de Haas, et al., 2021), and individual differences, for example based on children's existing word knowledge in their first language (van den Berghe et al., 2021b). The initial explorations culminated in a large-scale study, in which the combined use of a robot and a tablet was compared to using only a tablet to investigate the added value of having a robot physically present, and this study spanned multiple sessions to measure potential long-term effects (Vogt et al., 2019).

The L2TOR project played a leading role in the move within the human-robot interaction (HRI) field toward open science, as the project deliverables[2], source code[3], and data are made publicly available. Furthermore, to our knowledge, our study (Vogt et al., 2019) was the first publication at the HRI conference to have been preregistered, and several of our later studies (including Chapter 6) followed suit.

## 1.2 Gestures in education

Hand gestures can be described as 'visible actions' depicted with our bodies (Kendon, 2004). They serve an important communicative role, as they can be used to manage and guide the attention of others (Rohlfing et al., 2012), and can make information easier to understand compared to only using speech (Hostetter, 2011). Gestures can

[1] Additional information regarding the L2TOR project, including a promotional video, can be found on our website: https://web.archive.org/web/20210415022714/http://www.l2tor.eu/

[2] https://web.archive.org/web/20210415022714/http://www.l2tor.eu/

[3] https://github.com/l2tor

be beneficial in educational settings as well, where they are shown to aid students' understanding of the materials, particularly in difficult domains (e.g., language learning), or when dealing with complex or new concepts (Alibali & Nathan, 2007; Booth et al., 2008; Kelly et al., 2008; McGregor et al., 2009), potentially by reducing cognitive load (Goldin-Meadow, 2000). Students are likely to pay more attention to a lecture in which the teacher uses gestures (Valenzeno et al., 2003), which could lead to further improvements in their understanding and, hence, their learning outcomes.

Based on taxonomies from existing literature in gesture studies, we can distinguish between different types of gestures (Ekman & Friesen, 1969; McNeill, 1992). *Deictic* or pointing gestures are used to refer to objects, people, or locations, and thereby to direct someone's visual attention (Kita, 2003). *Beat* gestures can be used to emphasize certain parts of speech (Bosker & Peeters, 2021; Krahmer & Swerts, 2007). Meaningless motions that are subconsciously performed, such as self-touching, are referred to as *adaptors*. *Regulators* are gestures that facilitate various aspects of our communication, such as turn taking (Ekman & Friesen, 1969; Żywiczyński et al., 2017). *Emblematic* or *symbolic* gestures have a particular meaning that is agreed upon (e.g., waving to greet someone), and they can fulfil a social role or help structure the conversation (Kendon, 1995). *Iconic* and *metaphoric* gestures also convey meaning but, contrary to emblematic gestures, they elicit a mental image that automatically corresponds to the relevant concept, and as such their meaning does not have to be agreed upon (McNeill, 1985). Iconic gestures refer to concrete objects or actions, while metaphoric gestures refer to abstract concepts.

For iconic and metaphoric gestures to have a positive effect in education, it is argued that it is important for them to convey meaning, and that this meaning is understood by the student: Research has shown that arbitrary movements or gestures that do not match what is communicated via speech do not improve learning outcomes, while meaningful gestures do (Kelly et al., 2009; Macedonia et al., 2011). It is further proposed that these gestures are more effective if they are not completely redundant with what is communicated via speech, so that they add further details or examples (Hostetter, 2011).

There are different ways, also referred to as modes of representation (Müller, 2014), to depict a certain concept using iconic or metaphoric gestures. For example, the gesture for a *car* could consist of outlining the general shape of a car and its wheels, or the action of driving a car (or a combination of both). Although there is often a default gesture that most people would choose — in case of the car, likely the action of driving it — there is variation in how people transform their mental image of a concept into motion (Dargue & Sweller, 2018; Masson-Carro et al., 2017; Ortega & Özyürek, 2016; Ortega & Özyürek, 2020; van Nispen et al., 2014; van Nispen et al., 2017). This could be explained by the concept of *schemata*, proposed by Piaget and Cook (1952), referring to mental representations of objects and concepts that may differ between people. As we experience more aspects of an object or concept, our schema becomes more elaborate. In fact, observing gestures by others is said to facilitate this schematization process, by focusing the attention on particular salient aspects of a concept or event (Aussems & Kita, 2019). Variation theory (Marton & Booth, 2013) further states that an object of learning may be perceived differently between students, with one student's focus being on a different aspect of the object (e.g., the pointy shape of the tip of a pencil) than another student's (e.g., a pencil as a tool for drawing).

This variation in gesturing strategies can potentially also be explained by age: Children are more likely to use their entire body to depict a concept, while adults usually represent and manipulate the concept from an 'outsider looking in' perspective, by using only their hands (Sekine et al., 2018). For example, while children may form the tip of a pencil by raising both hands above their head to 'become' the object, adults will likely display the act of writing or drawing with an imaginary, but realistically sized pencil. Children generally also produce faster and less coordinated motions than adults (Jain et al., 2016), and the type of information they are trying to convey as well as their cognitive abilities have been shown to influence children's gesturing behavior (Abramov et al., 2021). Next to the mental image and age of the producer of the gestures, characteristics of the person the gestures are addressed to may play a role as well. Research on action demonstration has shown that adults perform different, often more exaggerated motions when they are addressing an infant compared to another adult (Rohlfing et al., 2006), and infants prefer these motions over adult-directed versions (Brand & Shallcross, 2008). This different way of demonstrating actions to infants is also referred to as 'motionese' — a variation of 'motherese' as infant-directed speech.

Not only do children produce gestures differently, they also appear to rely more on gestures performed by others than adults do (Hostetter, 2011). However, the ability to understand and make use of iconic and metaphoric gestures is a skill that develops during our early years (Novack et al., 2015; Stanfield et al., 2014), such that very young children (i.e., until three years old) may not be able to benefit fully from these gestures. There are further individual differences between children: When learning a second language, children with weaker skills in their first language appear to benefit more from gestures than those with stronger L1 skills (Rowe et al., 2013), which could be due to the fact that gestures are particularly useful for complex tasks (McNeil et al., 2000), and the task of learning words in a second language may be more difficult if the student is not as confident in their first language.

Preliminary research indicates that, in supporting language learning, gestures for particular types of concepts could be more effective than for other types of concepts. For example, one study observed that gestures referring to spatial concepts (e.g., *under*) or motor events (e.g., *running*) were more communicative than those referring to more abstract concepts (e.g., *blue*; Hostetter, 2011). However, students that were taught second language vocabulary in a different study still benefited from gestures even for abstract concepts (Repetto et al., 2017). Additionally, research has shown that gestures could be particularly useful for teaching verbs (Wakefield et al., 2018).

One final factor that may influence the effectiveness of gestures in education is whether the student also imitates or reenacts the gestures. Studies in solving mathematical problems (Cook et al., 2008), as well as first and second language vocabulary learning (de Nooijer et al., 2013; Repetto et al., 2017; Tellier, 2005, 2008) all showed that gestures were effective at improving learning outcomes when they were reenacted by the students, although the studies in second language learning did not compare between merely observing and reenacting the gestures. This beneficial effect of actively depicting the objects of learning might be explained by embodied cognition (Hostetter & Alibali, 2008) and, in the case of language learning, the language-action connection (Glenberg & Gallese, 2012).

To summarize, gestures are shown to have positive effects on (second language) education, although there are a number of factors that may influence their effectiveness: whether they are meaningful and understood by the student, how they relate to what is conveyed via speech, the chosen mode of representation, the age of the student, the type of concept that the gesture relates to, and whether the gestures are merely observed or also reenacted. Hence, the question arises whether these positive effects of gestures also apply when it is a robot performing them, instead of a human.

## 1.3 Robot-performed gestures

The previous sections have illustrated that gestures may be considered a key element of a robot's socially intelligent behavior, and an important way for a robot to make use of its physical presence. Because robots look human-like, they can perform similar gestures to those of human tutors, and these gestures may therefore also be effective at supporting the robot's tutoring efforts. However, currently available robots are limited in their motor degrees of freedom, which means that they cannot perform gestures with the same fluidity and detail as humans can. This, in turn, may have a negative effect on their performance, and thus it is important to evaluate whether the robot's gestures are elaborate and clear enough for them to be understood by the students.

Another important difference between gestures performed by humans and robots is that our gesture production process is spontaneous and subconscious (Hostetter & Alibali, 2008), while for robots the gestures will have to be designed and integrated with other modalities, such as speech. On the one hand, this gives the designer a large degree of control over the robot's behavior, so that gestures can be aligned with the robot's physical capabilities, and characteristics of the student (e.g., child-directed gesturing styles with exaggerated motions). On the other hand, this introduces a number of design decisions that may influence the effectiveness of the robot's gestures for better or for worse. For example, it is unclear what a desirable frequency of gesturing by robots is, but this is argued to be lower than a human's frequency to avoid forming a distraction (Pollmann et al., 2020).

Existing research into the effects of robot-performed gestures in education is scarce. However, studies in the field of storytelling indicate that a robot's use of gestures can help retain details of the stories (Huang & Mutlu, 2013; Szafir & Mutlu, 2012). In a learning-by-teaching scenario, where children had to correct the robot's reading, children with high reading proficiency themselves performed better if the robot used deictic gestures while reading, but children with low reading proficiency appeared to be distracted by the gestures (Yadollahi et al., 2018). To our knowledge, no previous studies, prior to the work presented in this thesis, have focused on second language learning. However, a study on memorizing first language verbs with adults showed that a robot's use of iconic gestures resulted in better recall compared to a robot that did not gesture (van Dijk et al., 2013). In addition, during interview studies, children of 10–12 years old (Ahmad et al., 2016a) and language teachers (Ahmad et al., 2016b) both mentioned seeing a potential in using robots as language tutors, and particularly stressed the important role that gestures could play.

Compared to robots that remain static, robots that gesture are often perceived as more human-like (Asselborn et al., 2017; Salem et al., 2013a), and are rated more positively (e.g., as likeable and lively; Huang & Mutlu, 2014; Salem et al., 2012). Furthermore, interactions with robots that gesture are found to be more enjoyable (Carter et al., 2014), and result in higher levels of engagement (Asselborn et al., 2017; Bremner et al., 2011), compared to robots that do not use gestures. These factors may, in turn, lead to the robot being perceived as a social agent (Bao & Cuijpers, 2017; Burns et al., 2018), and to the development of a relationship between the child and the robot (van Straten et al., 2020), which could further improve the effectiveness of the robot's tutoring efforts. Because the gestures are deliberately designed, it may be possible to exert some form of control over how the robot is positioned as a tutor, for example by providing it with a certain personality (e.g., based on Big Five traits, or in terms of dominance; Aly & Tapus, 2013; Peters et al., 2019), or an emotional state (e.g., positive or negative; J. Xu et al., 2014), which can help shape the way the robot is perceived by others.

In summary, based on existing research from related fields, robot-performed gestures show potential in supporting second language tutoring, as they could lead to improved learning outcomes as well as increased enjoyment and engagement, which could indirectly lead to better learning outcomes as well by means of a stronger social bond with, and greater acceptance of robots as tutors. However, these effects have not been empirically validated yet. There is also no clear overview of how different factors — relating to the design of the robot's gestures, their integration with the tutoring interaction, and individual differences between students — influence the potential beneficial effect of robot-performed gestures on second language learning. The current thesis aims to tackle these issues.

## 1.4 This thesis

In this thesis, we address the need for empirical research into the effects of robot-performed (iconic) gestures on second language tutoring, and identify factors that may influence these effects. Concretely, the following overarching research question will be addressed: *What are the effects of robot-performed gestures in the context of second language tutoring with children, and how are these influenced by the design* *decisions regarding the robot's gesture production process?* This research question can be divided into the following subquestions, which will be addressed in the upcoming chapters, and together shape our contribution to human-robot interaction and gesture studies:

RQ1 How can we best design and implement robot-performed iconic gestures? (Chapters 2–6)

RQ2 What are the observed benefits of robot-performed iconic gestures in human-robot interaction, and in robot-supported education in particular, according to existing literature? (Chapter 2)

RQ3 Does a robot that uses iconic gestures to support its second language tutoring efforts result in better learning outcomes than one that does not use iconic gestures? (Chapters 3, 4, and 6)

RQ4 Are children more engaged with a robot that uses iconic gestures, compared to with one that does not use gestures? (Chapters 3, and 6)

RQ5 What are potential factors that influence the effect of robot-performed iconic gestures on second language learning outcomes?
(Chapters 3, 4, and 6)

RQ6 How can we collect naturalistic human-performed examples of iconic gestures, and use these as input for designing a robot's gestures?
(Chapters 5, and 6)

RQ7 Do gestures contribute more to learning performance when multiple gestures are used for the same concept, highlighting different salient features of this concept, compared to a single gesture for each concept? (Chapter 6)

RQ8 Does variation in the robot's gesture repertoire result in higher levels of engagement with the robot or the task, compared to a single gesture for each concept? (Chapter 6)

## 1.5 Thesis outline

In Chapters 2–6, we present the results of applying a combination of methods in order to answer the research questions outlined above: a systematic literature review, three experimental studies that were conducted at primary schools in the Netherlands, and one dataset that was collected in a gameful human-robot interaction. This research was carried out as part of the Horizon2020 L2TOR project. For all studies except for the literature review, the SoftBank Robotics NAO V5 robot was used, which is a commercially available social robot with 25 degrees of freedom, shown in Figure 1.1.

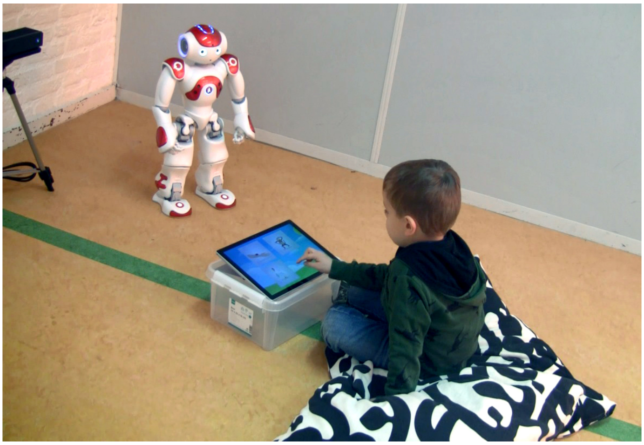


Figure 1.1: The NAO robot used in Chapters 3–6 of this thesis.

Chapters 2 and 4 are submitted to, and Chapter 5 is published in an academic journal. Chapters 3 and 6 were presented at the ACM/IEEE Human-Robot Interaction (HRI) conference, in 2018 and 2020 respectively. All chapters are therefore written as self-contained articles and, as a result, some overlap between the introductions (and with the general introduction), stylistic differences, and inconsistencies in the measurements and analyses used are unavoidable. However, the studies presented in these chapters do build upon each other to form a coherent narrative, and together contribute to answering the main research question.

**Chapter 2** Before addressing the main research question, it is important to create an overview of the state of the art regarding the design of a social robot's gesturing behavior, as well as the observed effects of these gestures. This was done by conducting a systematic review of existing literature in the field of human-robot interaction, with 167 articles that met the inclusion criteria. In this review, to create a comprehensive overview of existing research, we looked at robot-performed hand gestures in general, not only in the field of education. This allowed us to identify various factors that could potentially play a role in the effectiveness of the robot's gestures, as well as effects of gestures that could indirectly support a robot's tutoring efforts, such as changes in its perceived human-likeness.

**Chapter 3** In one of the first studies as part of the L2TOR project, also known as 'the animal experiment', we aimed to investigate whether a robot's use of iconic gestures could help Dutch children (4–6 years old) learn six animal names in English. This was done by having children play a game similar to *I spy with my little eye* with the robot, where the robot would 'spy' an English animal name, and the child then had to pick the corresponding image from a set displayed on a tablet device. In collaboration with our colleagues from Bielefeld University we also explored the role of adaptivity, where the order and frequency of presenting the animal names was based on children's performance on those words earlier in the interaction. Both aspects, gestures and adaptivity, were combined in one study because they could both affect the level of difficulty for the child, and thus influence learning outcomes and levels of engagement.

**Chapter 4** Based on promising findings from the study described in Chapter 3, the robot's use of iconic gestures was further investigated as part of a more complex, longitudinal study, spanning seven sessions. The tutoring interaction in general was now more complicated than the game of *I spy with my little eye*, where children and the robot visited different three-dimensional virtual environments on the tablet, interacted with objects in this environment, repeated English words, and enacted a number of verbs. The 34 English words included in this study were also more complex than the animal names from the study in Chapter 3, for example including prepositions such as *next to*. The general results of this study, which were published at the HRI conference (Vogt et al., 2019), showed no benefit of the robot's use of gestures on children's learning outcomes. In Chapter 4, we present additional analyses from this experiment to investigate whether several factors — the clarity and 'quality' of the design of the gestures, differences between types of English words, age of the children, and spontaneous reenactment of the gestures — had an influence on the observed learning outcomes.

**Chapter 5** One of the main insights from the literature review (Chapter 2) is that the process of designing the robot's gestures is often not given much thought, and that gestures are commonly based on a researcher's idea of what they should look like. At the same time, literature in gesture studies indicates that people might use different ways to represent a concept via gesture, and it is conceivable that children perform — and perhaps also expect to see — different gesture forms than adults. While existing datasets of human-performed gestures do exist, most of them are collected in a lab, where participants were constantly aware that they were providing these data, and they were given concrete prompts, e.g., to perform *brushing your teeth* rather than *toothbrush*. This does not capture any variation in gesturing approaches that is likely to occur in a naturalistic environment. In Chapter 5, we therefore present a dataset of human-performed gestures, collected from children and adults in the field: at the NEMO science museum and the Lowlands music festival, both in the Netherlands. Participants engaged in a game of charades with the robot, to ensure that they were not constantly aware of being recorded. The dataset can be used to base the design of a robot's gestures on real-world examples, and it also addresses a need in the field of gesture studies for more data that can be used to study the human gesturing process (i.e., variation, links between kinematics and semantics, differences between children and adults).

**Chapter 6** To address the mixed findings from Chapters 3 and 4, in this study we revisited the relatively simple game of *I spy with my little eye*, but now including animal names as well as more complex and abstract English words, such as *bridge*. The robot's gestures were inspired by recordings from the dataset introduced in Chapter 5, but they were recreated by hand to take into account the robot's physical limitations, and to ensure that they were as clear as possible. From Chapter 5 we observed that, indeed, there is variation in how people depict concepts using gestures. Initial explorations in educational sciences further indicate that providing variation in stimuli (i.e., using multiple images or different voices to train vocabulary) could lead to better learning outcomes. Therefore, the study documented in Chapter 6 contains an experimental condition in which the robot used five different gestures for each English word, rather than repeating the same gesture every time this word was trained. We measured whether the robot's use of iconic gestures, either repeated or varied, affected children's learning outcomes and their levels of engagement with the task and the robot, and whether age played a role.

**Chapter 7** In this final chapter, we reflect upon the previous chapters and integrate their findings, in order to answer the eight research questions, and finally the main research question of this thesis. We also propose avenues for future work, and introduce our vision of the future of using social robots for education.

2

Chapter 2

# Robot-Performed Gestures: A Systematic Review

*This chapter is based on:* de Wit, J., Vogt, P. & Krahmer, E. The design and observed effects of robot-performed manual gestures: A systematic review. Submitted for journal publication. *Open practices:* The data are available at https://osf.io/uj9fq/

## Abstract

Communication using manual (hand) gestures is considered a defining property of social robots, and their physical embodiment and presence, therefore we see a need for a comprehensive overview of the state of the art in social robots that use gestures. This systematic literature review aims to address this need by (1) describing the gesture production process of a social robot, including the design and planning steps, and (2) providing a survey of the effects of robot-performed gestures on human-robot interactions in a multitude of domains. We identify patterns and themes from the existing body of literature, resulting in ten outstanding questions for research on robot-performed gestures regarding: developments in sensor technology and AI, structuring the gesture design and evaluation process, the relationship between physical appearance and gestures, the effects of planning on the overall interaction, standardizing measurements of gesture 'quality', integrating gestures with other modalities, individual differences, gesture mirroring, whether human-likeness is desirable, and universal accessibility of robots. We also reflect on current methodological practices in studies of robot-performed gestures, and suggest improvements regarding replicability, external validity, measurement instruments used, and connections with other disciplines. These outstanding questions and methodological suggestions can guide future work in this field of research.

## 2.1 Introduction

**Social robots** are designed to engage in everyday interaction with human interlocutors, such as visitors to a store or students in a classroom. Combined with the use of natural language to communicate, robots can leverage their embodiment and their presence in the physical world to exhibit socially intelligent behavior (Bartneck & Forlizzi, 2004; Duffy & Joue, 2000). An essential part of this socially intelligent behavior is facilitated by the embodied nature of robots and consists of non-verbal communication, such as gaze and **manual (hand) gestures** (Fong et al., 2003). While a comprehensive overview of social gaze behavior by robots and its effect on human-robot interactions can be found in HRI literature (Admoni & Scassellati, 2017), to the best of our knowledge a survey of the state of the art and open research questions for robot-performed manual gestures is still missing. We see a need for such a survey, because robot-performed gestures are often highlighted as a defining property of social robots in existing literature studies (e.g., Johanson et al., 2020; Mavridis, 2015; Robert et al., 2020; Saunderson & Nejat, 2019; Skantze, 2020), but are typically not further discussed in detail. In addition, a robot's ability to use gestures within the physical world has been shown to be one of its main advantages over alternatives, such as virtual agents or robots presented on a screen (Li, 2015), which can therefore be considered one of the main channels through which social robots can make use of their physical embodiment and presence. With the increasing interest in social robots from research and industry across a multitude of domains, including the hospitality industry (Ivanov et al., 2017), healthcare (Cifuentes et al., 2020), and education (Belpaeme et al., 2018), it is important to optimize the design and application of the robot's gestures. In this chapter, we aim to address this need for a comprehensive overview of robot-performed gestures by presenting the results of a **structured and exhaustive literature study** into (1) the design and planning approaches to robot-performed gestures, and (2) the effects that these gestures have on human-robot interactions. We address these two topics together, because we consider them to be strongly related: In order for a robot's gestures to have an effect, they will need to be well designed and implemented.

### 2.1.1 Robot versus human gestures

Because gestures of social robots tend to look human-like, we can derive conjectures about the potential effects of robot-performed gestures from observations regarding human produced gestures. Manual (hand) gestures are seen as an important element of our communication with others, as they help us to convey our intentions, attitudes, feelings, and ideas — either voluntarily or involuntarily (Kendon, 2004). McNeill (1992) distinguishes between four **types of gestures**: *deictic* or *pointing gestures* that help guide someone's attention toward a particular object, location, or person; *beat gestures* that can be used to emphasize parts of speech; *iconic gestures* that depict a particular action or object (e.g., molding the shape of a ball with our hands); and *metaphoric gestures* that relate to abstract concepts (e.g., time or size). Iconic gestures can be performed in different ways, for example by portraying actions (with or without involving imaginary objects), or movements that outline shapes of objects (Müller, 2014). Additional gesture types, such as *emblems* (Beattie, 2003; Ekman & Friesen, 1969), also known as *symbolic gestures* (Krauss & Chen, 2000), have further been proposed. These refer to gestures for which the meaning has been agreed upon, for example waving as a form of greeting someone. Finally, *adaptors* have been defined both as meaningless idle motions, such as self-touching (Ekman & Friesen, 1969), as well as motions that are used for managing the flow of conversation, in that case also called *regulators* (Ekman & Friesen, 1969). Although most gesture types can be used in isolation, which is referred to as pantomime (McNeill, 1992), they are often integrated with speech. In the latter case, the gestures can be redundant, in that they convey the same message that is communicated through speech, or they can be complementary by adding information to the co-occurring verbal utterance (Goldin-Meadow, 2005; McNeill, 1992), for example by waving in the far distance while mentioning a particular city. This avoids the listener having to recall the distance between someone's home and the city they travelled to.

Research has shown a number of **beneficial effects from the use of these different types of gestures**. Iconic and metaphoric gestures have been shown to facilitate communication, by aiding the speaker's language production process (Cravotta et al., 2019; Hostetter, 2011) and by making it easier for the listener to comprehend the multimodal message that is communicated to them (Arachchige et al., 2021; Goldin-Meadow, 2005; Hostetter, 2011; Kelly et al., 1999). Because these gesture types convey meaning, they also play a supportive role in education by improving the memorization of new, complex concepts (Alibali & Nathan, 2007; Kelly et al., 2008), and students tend to be more engaged with a teacher that uses gestures compared to one that is static (Valenzeno et al., 2003). For iconic gestures to be beneficial to learning and to communication in general, it is important that the gestures are congruent with the concept that is conveyed via speech (Kelly et al., 2009; Kelly et al., 2010b; Macedonia et al., 2011). Deictic gestures play an integral role in early language acquisition (Iverson & Goldin-Meadow, 2005), and in facilitating joint visual attention (Kita, 2003; Louwerse & Bangerter, 2005; Tomasello et al., 2007). Beat gestures have been shown to provide stronger emphasis on specific spoken words compared to using only prosody (Bosker & Peeters, 2021; Krahmer & Swerts, 2007), while regulators and adaptors can help structure conversations, for example by creating a more natural turn taking process (Mlakar et al., 2021; Żywiczyński et al., 2017). Finally, emblems often play a social role by conveying a certain emotional state or politeness (Kendon, 1981, as cited in Lindenberg et al., 2012), and can also be used to structure the discourse, or to communicate illocutionary intent — information that is 'between the lines', e.g., conveying that what is said as a statement is actually meant to be a request (Kendon, 1995; Kita, 2009).

As humans, our gesturing behavior is to a large extent automatic and spontaneous (Hostetter & Alibali, 2008): we do not consciously plan the type of gesture, shape, or timing while we are speaking. For robots, however, the **gestures will have to be shaped or designed**, and the correct way to perform them will have to be implemented. This includes selecting a gesture to depict, and determining the timing of this gesture, for example to coincide with speech or to fit the current (social or emotional) context. How these gestures are designed and integrated may affect how the robot is perceived, and to what extent the robot manages to successfully support humans in the tasks they are trying to perform. In addition, robots that are currently available are more limited in their gesturing capabilities, in terms of degrees of freedom and fluidity of movement, compared to humans. This raises the question whether robot-performed gestures can provide the same positive effects that are observed with human-performed gestures. Although there is a large number of studies addressing this question, a comprehensive overview is currently lacking.

### 2.1.2 Why this literature review

Robots are not able to gesture automatically without first implementing the **design and planning** steps of the gesture production process. Several design decisions will have to be made in order to create gesturing behavior that achieves the desired positive outcomes in human-robot communication. For example, it is important that the robot's gesture and speech are correctly aligned (Mavridis, 2015). While literature reviews do exist on specific topics related to gesture design, including the use of animation techniques (Schulz et al., 2019), or motion design for a broad range of robot platforms including non-humanoids (Venture & Kulić, 2019), the subsequent planning steps including (1) gesture selection, (2) co-speech timing, and (3) adjusting motion parameters based on contextual information are not covered in detail in these existing reviews. We believe there is a need for unified, clear guidelines for the creation of robot-performed gestures, and therefore the first aim of this literature review is *to provide a survey of ways in which the design and planning steps of a robot's gesture production process can be implemented.*

Although the **effects of robot-performed gestures** have emerged as a theme in existing literature reviews — for example in the aforementioned comparison with telepresent robots and virtual agents (Li, 2015), as well as surveys on robot personalities (Robert et al., 2020) and turn taking (Skantze, 2020) — to our knowledge there is no review that focuses specifically on the effects of the robot's use of manual gestures in human-robot interactions. In addition, several existing reviews take on a holistic approach when studying robots' communicative behaviors, including verbal components (Johanson et al., 2020; Mavridis, 2015), or other non-verbal channels, such as proxemics and haptics (Saunderson & Nejat, 2019). Because of the broader nature of these reviews, they are only able to cover the tip of the iceberg regarding humanoid robots and their use of manual gestures. Even though in recent years various surveys have appeared, focusing on the state of the art in social robotics, for example for educational purposes (Belpaeme et al., 2018; Kanero et al., 2018b; van den Berghe et al., 2019), they do not address the role of gestures, even though the ability to produce manual gestures is recognised as one of the core properties of social robots. With the increasing use of humanoid, social robots in a wide range of application domains, there is a need for a comprehensive overview of the effects that manual gestures have on the quality of the communication between humans and robots, specifically in terms of (1) communicative purposes, including behavioral responses (e.g., mimicry), (2) the perception of the robot (referred to as 'cognitive framing' in Saunderson & Nejat, 2019), (3) engagement, and (4) the performance of the robot and human interlocutors on various tasks, such as learning or direction-giving. The second aim of the present literature review is therefore to *compile a comprehensive overview of the effects of robot-performed gestures in the multitude of domains where social robots could be used.* We believe that these effects rely on the approaches taken regarding the design and planning of the robot's gesturing behavior, which is why both aspects — the gesture production process, and the observed effects — will be addressed together in this chapter.

## 2.2 Methodology

A **systematic search** of published scientific articles was conducted in order to identify relevant sources for the two main objectives of the present literature study: (1) a survey of implementations of the design and planning steps, and (2) a comprehensive overview of the effects of robot-performed gestures. This review was conducted following the updated Preferred Reporting Items for Systematic reviews and Meta-Analyses guidelines (PRISMA 2020; Page et al., 2021a; Page et al., 2021b).

### 2.2.1 Data sources and search strategy

In order to find relevant literature, the following **broad search queries** were used:

- robot AND gestures AND (humanoid OR social)
- robot AND iconic AND gestures
- robot AND pantomime

We did not use literal searches (with quotation marks) so that synonyms or related terms (e.g., singular 'gesture') would also be included. To get a comprehensive overview, the search queries were run on **three different databases that are relevant in the field of human-robot interaction**:

- Association for Computing Machinery (ACM) Digital Library
- Institute of Electrical and Electronics Engineers (IEEE) Xplore
- Web of Science

These databases work with different formats and have different interfaces. In the ACM DL, all fields including the full-text were searched because searching only abstract and title appeared to return limited results and this required manually editing the query syntax. For IEEE Xplore and Web of Science all metadata were included but not the full-text. Data were collected in March, 2021. Additionally, during our search, a limited number of additional papers ($n = 8$) came to our attention, either while examining related work of articles that were already included in the study, suggested by the authors of previously included articles, or serendipitously while working on other publications and reviews. These were included as well.

### 2.2.2 Selection process

The results of running the search queries on the three databases were exported to the BibTeX format, after which they were loaded into the Mendeley reference manager[1] for conducting the selection process. This process consisted of two stages: **initial screening** based on title and abstract, followed by **full-text assessment** for eligibility. The following inclusion criteria were taken into account:

- The article documents either a method for planning robot gesturing behavior (objective 1), OR an experimental study in which the effects of robot-performed gestures were studied (objective 2);
- A humanoid robot was used. This excludes virtual agents, and non-humanoid robots, such as industrial or zoomorphic (animal-like) robots. In some cases, articles involving a virtual simulator were included, but only if this was a simulator of an existing humanoid robot platform;
- The robot performed communicative manual (hand) gestures;
- The study focused on robot-performed gestures, not recognition of human-performed gestures by robots;
- The robot is designed to interact with humans;
- The article is published in a scientific journal or conference proceedings;
- The article is written in English.

These criteria were used when assessing full-texts for eligibility, and in most cases could already be evaluated when screening abstracts and titles. If the type of robot used was unclear during this initial screening step, the full-text was briefly scanned for images of the robot. Articles documenting the design of robot-performed gestures were not explicitly selected and included, as this design step tends to be a prerequisite for gesture planning implementations and for experimental studies. Therefore, we inferred the different design approaches from the papers already included in the dataset based on the inclusion criteria, and searched for additional literature explaining these approaches when needed.

[1] https://www.mendeley.com/

After careful discussion of the criteria by all authors, the selection process was conducted by one researcher, who took on a lenient approach in the initial scanning stage by including papers for which it was unclear whether they met all criteria. Because the inclusion criteria were relatively unambiguous and there were no doubts about the eligibility of any of the included articles after assessing the full-text, it was not deemed necessary to include a second reviewer. The selection process, including the number of articles included in each stage, is displayed in Figure 2.1.

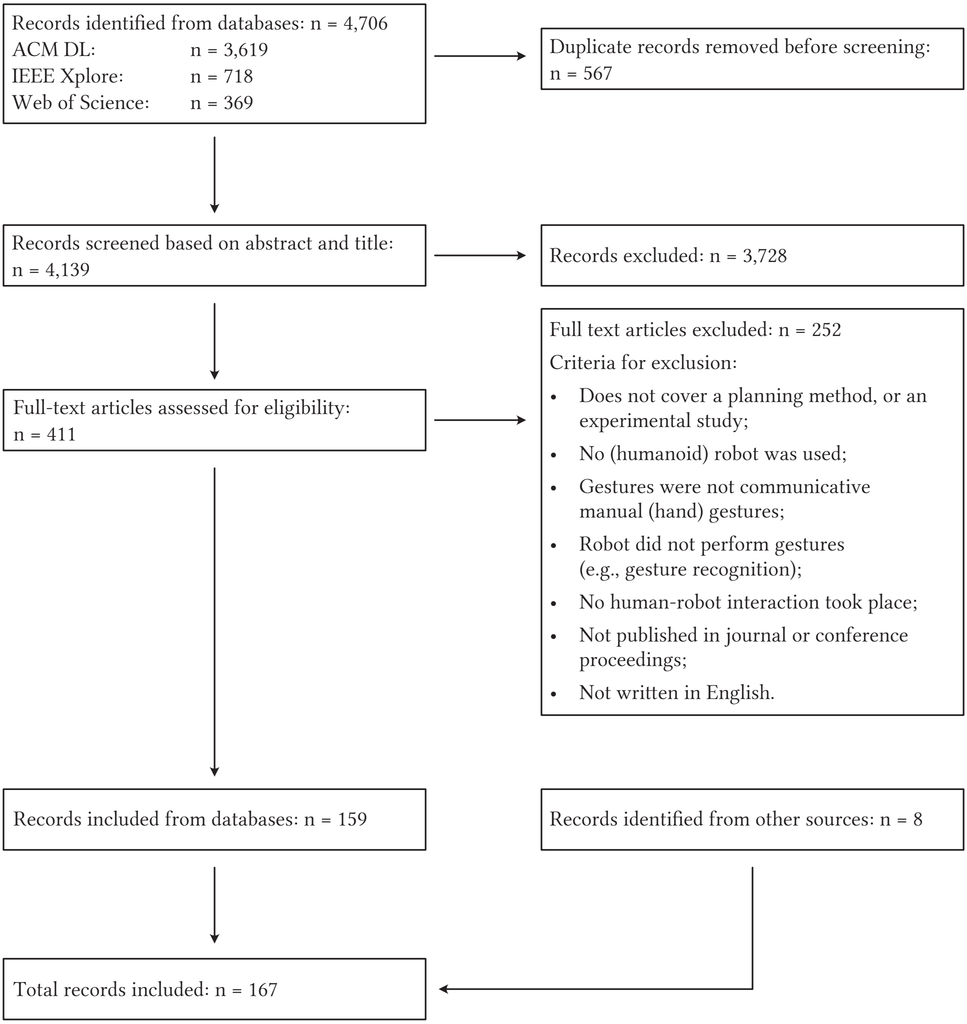

Records identified from databases: n = 4,706
ACM DL: n = 3,619
IEEE Xplore: n = 718
Web of Science: n = 369
Duplicate records removed before screening: n = 567
Records screened based on abstract and title: n = 4,139
Records excluded: n = 3,728
Full-text articles assessed for eligibility: n = 411
Full text articles excluded: n = 252
Criteria for exclusion:
• Does not cover a planning method, or an experimental study;
• No (humanoid) robot was used;
• Gestures were not communicative manual (hand) gestures;
• Robot did not perform gestures (e.g., gesture recognition);
• No human-robot interaction took place;
• Not published in journal or conference proceedings;
• Not written in English.
Records included from databases: n = 159
Records identified from other sources: n = 8
Total records included: n = 167

Figure 2.1: Overview of the article selection process.

### 2.2.3 Information obtained from each article

The **167 articles** that were included in the review after the selection process were labeled as describing gesture planning, an experimental study, or both. From all included papers, the following information was obtained:

- Year of publication
- Robot platform used (see Appendix 2.A for images)
- Context or application domain
- Region in which the research took place
- Seminal gesture theory references that were cited
- Gesture type, and examples of gestures
- Gesture design approach, who designed the gestures

In addition, for papers documenting a planning approach we obtained:

- The goal of the planning approach (e.g., gesture selection, co-speech timing)
- The method or technical implementation that was used (e.g., neural network)
- Whether and how the proposed system was evaluated

From papers that presented an experimental study we obtained:

- Research question
- Sample size and demographic
- Location of the study (lab, field, online)
- Number of sessions with the robot
- Planning approach that was used
- Experimental design
- Whether and how the quality of the design of the gestures was assessed (e.g., comprehensibility)

- Effects studied and measurement instruments used
- Findings
- Suggestions for future work

Finally, papers were assigned a **theme** (e.g., ‘communicative purposes’), which allowed us to structure the following results sections, and we noted whether the effects found in experimental studies were positive, negative, neutral, or mixed, or if no effects were found. The tables summarizing the included articles can be found in Appendix 2.B (planning) and 2.C (experimental studies), and the full versions of the tables containing all of the information mentioned above are made available as supplementary materials[2].

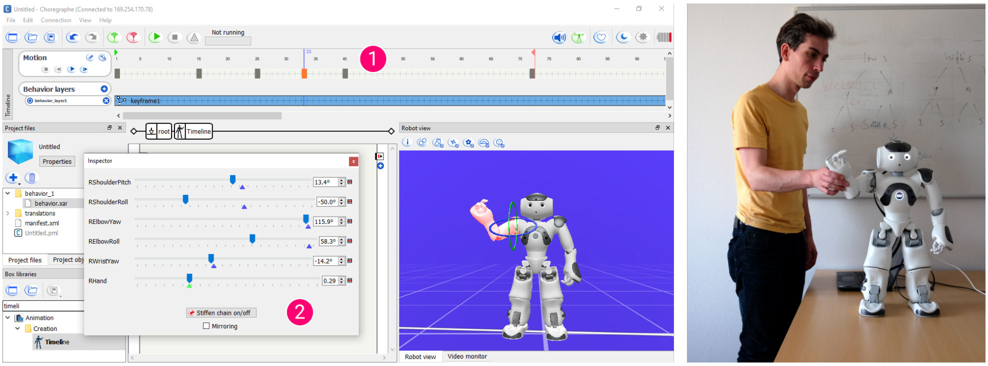
## 2.3 Gesture design

Social robots are used in areas where they are intended to interact with humans, providing services, such as teaching, that we are used to receiving from other people. Therefore, generally the aim is to also have the robot use gestures that we know from our communication with other people, because these match the expectations we have of how to communicate non-verbally. However, because there are differences in the freedom and fluidity of movement, both between robots and humans and between the different robot platforms that are available, it is often not possible to directly transfer human-performed motion onto the robots. Consequently, these gestures will have to be designed, oftentimes specifically for one robot platform. There are two ways to approach this design process: manually, or from demonstration.

### 2.3.1 Manual gesture design

Manual design means that someone — e.g., the researchers, or a professional animator — implements a gesture step-by-step. This is generally done by **defining a series of key frames**, representing salient points within a motion (area 1 in Figure 2.2). By moving the key frames around on the timeline a movement can be either sped up or slowed down, and by having the same frame twice, the robot can hold a pose for a desired amount of time. The robot then performs this sequence of key frames as gestures, by smoothly interpolating between the defined poses at a set speed. This technique has its origin in the field of character animation, and is therefore also used

[2] https://osf.io/uj9fq/



RShoulderPitch 13.4°
RShoulderRoll -50.0°
RElbowYaw 115.9°
RElbowRoll 58.3°
RWristYaw -14.2°
RHand 0.29
□ Mirroring

Figure 2.2: Manual gesture design by defining key frames (left, area 1), either by adjusting the values of the robot's joint orientations (left, area 2) or by physically moving the robot's limbs into the desired positions (right). This example uses the Choregraphe interface by SoftBank Robotics (Pot et al., 2009); different interfaces exist for other robots.

in contexts other than human-robot interaction, such as cartoons or games. The key poses can be defined by setting the coordinates or angles of the robot's joints using an interface or code (area 2 on the left side of Figure 2.2), or by physically moving the robot's limbs to the correct position and then storing the joints' information, a process known as kinesthetic teaching or puppeteering (Figure 2.2, right picture).

The main advantage of the manual approach to designing gestures is that the designer has **full control over what the gestures will end up looking like**. The designer can draw inspiration from human-performed gestures, for example by recreating gestures that have been observed in field studies or by conducting an elicitation study, and they can also use animation techniques to add more expressiveness and exaggeration (Marmpena et al., 2019) (for an overview of the use of animation techniques in human-robot interaction scenarios, see Schulz et al., 2019). At the same time, because the process is done by hand it is possible to work around the particular robot's physical limitations, in order to create gestures that could be considered optimal given the current robot platform's capabilities. The main disadvantage of this approach is that it tends to be labor intensive as each gesture needs to be defined step-by-step, although parts can often be reused. In addition, the gestures may appear artificial and predefined because of the smooth interpolation between key points, so there is no degree of variation or human-like noisy movement included in the motions.

### 2.3.2 Gesture design from demonstration

Manually designed gestures can be inspired by human-performed gestures, by observing recorded gestures and then recreating them using key frames to accommodate the robot's physical properties. It is also possible to record human motion and then translate this automatically onto the robot, while taking into account differences in morphology. This process is known as learning from demonstration or imitation learning, and was originally used to teach robots how to perform certain tasks, such as manipulating objects in the physical world (Argall et al., 2009). A real-time **mapping of human movement** is also applied for telepresence purposes, where the robot is used as a physically present representation of someone who is presenting or attending a meeting remotely (e.g., Bremner & Leonards, 2015a). Traditionally, human motion is recorded using motion capture technologies, complex multi-camera set-ups that often require the performer to wear markers. However, recently more compact depth sensors (e.g., Microsoft's Kinect) have become available, which allow markerless tracking with only one portable camera, although at the cost of recording accuracy (Figure 2.3). Even more recently, using advanced AI and computer vision techniques researchers have managed to extract three-dimensional motion recordings from data found 'in the wild', such as YouTube videos (e.g., Hua et al., 2019; Shimazu et al., 2018; Yoon et al., 2019; Yu & Tapus, 2020). After collecting recordings of human-performed gestures, a form of mapping or translation of these recordings is required, because there are differences in both size and kinematic abilities between the original human performer and the robot that is copying the gesture — this is known as the *correspondence problem* (Mohammad & Nishida, 2013). This mapping is done either algorithmically, e.g., by calculating the angles between various joint locations, or using neural network-based approaches (e.g., Matsui et al., 2005).

Compared to manual gesture design, learning from demonstration is **less labor intensive**: Once the mapping is in place any number of recordings can be transferred onto the robot. This also makes it easier to introduce variation, and the gestures are more detailed because they are defined frame-by-frame, instead of interpolating between key frames (Marmpena et al., 2019). Additionally, aspects such as intention and enthusiasm, that are subconsciously incorporated into human gesturing behavior, can be included when directly transferring recordings onto the robot (Shimazu et al., 2018). The main drawback to the method is that it is difficult to take the differences between the human and the robot, in terms of their physical appearance (e.g., height) and kinematic abilities, into account. This invariably results in a loss of detail when

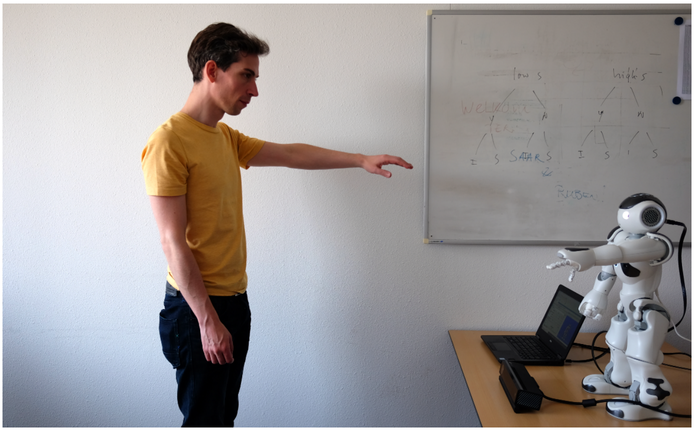


Figure 2.3: Example of a real-time learning by demonstration set-up, with a human performer that is recorded by a depth camera, and a robot that is directly copying the human's movements. Note: the Microsoft Kinect is positioned incorrectly, for it to fit in the photo.

translating the gestures from the original recording onto the robot (Venture & Kulić, 2019). In addition, the resulting robot-performed motions can contain movement or 'jerkiness' that was not present in, or different from the original gesture because of imperfect capturing by the sensor or inaccuracies in the mapping process. This can however be partly remedied by applying post-processing steps, such as denoising and dimensionality reduction (e.g., Kucherenko et al., 2019) to create a smoother gesture.

### 2.3.3 Platform-agnostic gestures

There is a wide variety of humanoid robot platforms available, and gestures are often developed specifically for one particular robot, in order to take optimal advantage of its physical characteristics. To facilitate replicability of studies, and to allow for a fair comparison between different robots, it would be valuable to have a platform-agnostic gesture representation. For manual design approaches, oftentimes proprietary software and methods are used that are unique to a particular robot platform, making it difficult to create gestures that can easily be transferred onto a different robot. When applying learning from demonstration approaches, the recordings of human gestures — particularly after reducing their complexity (Kucherenko et al., 2019) — could be seen as such a platform-agnostic representation, which can then be mapped to different robot platforms. However, as previously described these mappings still cause a substantial loss of information to occur. Therefore, several higher-level, more **abstract motion representations** have been proposed, such as the Laban notation (Von Laban, 1975) which was originally designed to describe dance movements, as well as descriptions of basic, small motions (e.g., 'raise right hand above head') that are implemented on multiple platforms and can be used as 'building blocks' to create complex gestures (van de Perre et al., 2018). However, the vast majority of the research covered in this review focuses on a single robot platform, and only implements gestures for that particular platform.

## 2.4 Gesture planning

After deciding on a gesture design strategy, the second step in providing robots with the ability to produce gestures is planning. At this stage, a relevant gesture has to be picked from the available lexicon of gestures, and it has to be integrated with the other communicative modalities, for example speech. Optionally, in this step the robot could make adjustments to the gestures depending on various factors, such as the emotional state that the robot wants to convey, or the physical location of objects or people that the robot wants to refer to. A total of **61 papers** included in this study present a gesture planning approach, and will therefore be discussed in the upcoming sections.

### 2.4.1 Methodological characteristics of planning approaches

We observed the following characteristics when inspecting the set of 61 papers that introduce a planning technique:

- **Focus/theme:** 18 papers cover gesture selection (Section 2.4.2), 6 co-speech timing (Section 2.4.3), 12 gesture synthesis (Section 2.4.4), and 37 adjusting motion parameters (Section 2.4.5);

- **Gesture types produced by the robot:** 36 emblematic (symbolic, meaning is agreed upon), 30 deictic (pointing), 21 metaphoric, 20 iconic, 18 beat, 9 adaptors (1 unknown);

- **Gesture theory references** 36 out of 61 papers (59%) explicitly cite studies on human gesture production;

- **Commonly used robots:** NAO (22), Pepper (13), Erica (4), ASIMO (2), iCub (2), Wakamaru (2);

- **Gesture design approach:** 30 manual, 17 from demonstration (15 unknown);
- **An evaluation of the planning approach** is presented in 48 out of 61 papers (79%).

A table summarizing key details of the papers covered in this chapter is included in Appendix 2.B, and the full table is available in the supplementary materials[3].

### Evaluating gesture planning

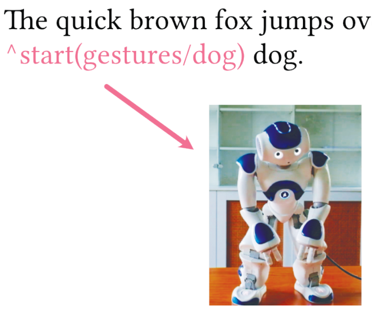
The quality of the different planning approaches (selection, timing, synthesis, and adjusting the motion parameters) can be measured in two different ways. First, this can be done by using computational methods, for example by comparing an algorithm's accuracy at selecting and timing gestures to a ground truth from a dataset or by measuring gesture traits, such as 'jerkiness'. Second, the quality can be assessed using evaluation studies with human subjects, by means of a live study or by recording an interaction and embedding this into a survey. A subset of these evaluation studies incorporates the gesture planning approach into an experiment where the effects of the robot's gesturing behavior are also studied — these are studies that also appear in Section 2.5.

Of the 61 papers included in this part of the review, 16 (26%) used computational methods, 36 (59%) included an evaluation of the proposed planning method with human subjects, and 13 (21%) did not evaluate their approach. Four papers (7%) used a combination of computational methods and human evaluation studies, which is why the total amounts to more than 100%. Of the 36 human evaluation studies, 19 (53%) studied the effects of the robot's gesturing behavior on the resulting human-robot interaction and are therefore also included in Section 2.5. The sample size of the evaluation studies with human subjects ranged from 10 to 396 ($M = 46, Mdn = 30, SD = 65$; 1 unknown). These studies mainly focused on assessing aspects such as the naturalness of the robot's behavior, or whether participants were able to understand the message or emotion that the robot was conveying, and this was commonly done using self-report measures.

## 2.4.2 Gesture selection

As previously mentioned, human gesturing generally occurs spontaneously and automatically, and typically while speaking (Hostetter & Alibali, 2008). This has to

[3] https://osf.io/uj9fq/



The quick brown fox jumps over the lazy ^start(gestures/dog) dog.

Figure 2.4: Gesture selection by manually annotating the script (example from the markup language used in the Choregraphe tool for controlling the NAO robot; Pot et al., 2009). The robot performs the gesture for ‘dog’ when the annotated tag (highlighted in the text) is encountered in its speech output.

be programmed deliberately for robots, which can broadly be done in two ways: by manually annotating the script of the robot’s speech output to indicate when a particular gesture should be performed, or by mapping between words and gestures so that a gesture is triggered whenever that word occurs in the robot’s speech. The latter method does not take various contextual factors, such as the preceding words or gestures into account, which can be modelled separately and then included in the gesture selection process.

The **manual annotation** approach is commonly done by adding tags to the robot’s utterances (Figure 2.4), describing which gestures to perform, and when to perform them (Ondáš et al., 2017; Shi et al., 2010). This method offers the designer full control over the robot’s behavior, although it is also the most labor intensive approach and it requires the interaction designer to plan all of the robot’s utterances ahead of time. Several markup languages exist to simplify and structure the process of defining multimodal output for robots and virtual agents, such as:

- Multimodal Utterance Representation Markup Language (MURML; Kranstedt et al., 2002);
- Behavior Markup Language (BML; Vilhjálmsson et al., 2007);
- Function Markup Language (FML; Heylen et al., 2008).

To make gesture selection more scalable and less labor intensive, several researchers have started creating **mappings between words and corresponding**

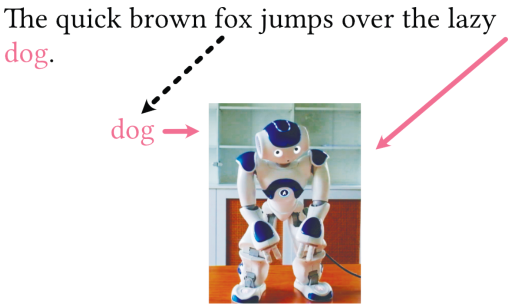


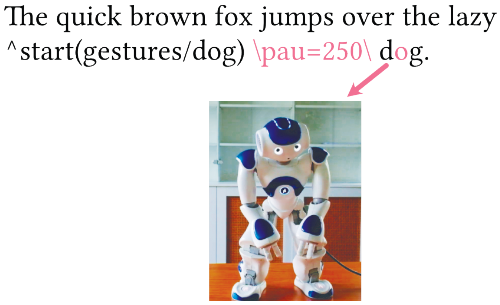
Figure 2.5: Gesture selection by triggering a gesture whenever a particular word is encountered. In this example, the robot performs the gesture for ‘dog’ whenever the word ‘dog’ occurs in its speech output. Using lexical databases, the robot can also trigger the gesture for semantically similar words, such as ‘fox’, if no gesture for ‘fox’ is available.

**gestures**. Whenever such a word is encountered in the robot’s speech, the robot will simultaneously perform the matching gesture (Ondáš et al., 2017). These mappings can be manually defined, or inferred from data of human speech and gesturing (Ghosh et al., 2019). By using lexical databases, such as WordNet (Miller, 1995), it is possible to make this system more intelligent by also linking gestures to words that are semantically similar to the originally assigned word (Augello and Pilato, 2019; Figure 2.5). This approach prevents the designer from having to predefine all of the robot’s utterances or determine all of the mappings, and therefore supports more dynamic natural language generation. Initial explorations in human gesturing behavior indicate that the gestures of words that are semantically similar share kinematic similarity as well (Pouw et al., 2021), a further indication that using lexical databases could be a cost-effective way to expand the robot’s gesturing behavior. A commonly used system to support this process of linking words and gestures is BEAT (Cassell et al., 2004), which was originally created to generate multimodal output for virtual agents and has since also been applied in robotics. Researchers have also begun exploring neural networks as a way to automatically link gestures to generated utterances based on large datasets (Hwang et al., 2020).

A drawback of these automated methods is that they do not **take context into account**. For example, beat gestures have been found to lead to more engagement from the interlocutor if they are performed in a way that naturally connects with the preceding gesture (Bremner et al., 2009). The robot should also not gesture too frequently to avoid confusing or frustrating others, and the situation could arise that the automated gesture selection method proposes multiple gestures, such as an iconic gesture and a beat gesture, at the same time. Predictive and probabilistic models enable the robot to include additional sources of information other than its speech output when deciding whether to gesture, and if so, what type of gesture to perform. Examples of factors that are included in these decisions are prosody and linguistic features (Ishi et al., 2018; Mlakar et al., 2013; Pérez-Mayos et al., 2020), the desired level of expressiveness of the gestures (Ng-Thow-Hing et al., 2010; Tay & Veloso, 2012), priority of a certain motion over another (Sunardi & Perkowski, 2020), the affective state of the robot (Paplu et al., 2020) or its audience (Aly & Tapus, 2020; Bourguet et al., 2020a), or the current dialog and environmental context (Admoni et al., 2016; Augustine et al., 2020; Ishi et al., 2018; S. Lim et al., 2009). The logic behind these models can be based on rules and probabilities that are learned from annotated recordings of human-performed multimodal communication (Huang & Mutlu, 2014; Ng-Thow-Hing et al., 2010). The fact that the robot shows greater intelligence regarding its gesturing behavior with the use of these models has been shown to result in interactions that are perceived as more natural compared to implementations that only take into account words as triggers (Ishi et al., 2018).

### 2.4.3 Co-speech timing

After determining which gesture to perform, the motions also need to be timed in such a way that they correspond to the related information that is conveyed through the robot's speech. This can be done, for example, using a process of trial-and-error



```
The quick brown fox jumps over the lazy
^start(gestures/dog) \pau=250\ dog.
```

Figure 2.6: Manually annotating the script to control gesture timing (example from the markup language used in the Choregraphe tool for controlling the NAO robot; Pot et al., 2009).

to find a starting point that allows the *stroke* of the gesture to coincide with the corresponding word (e.g., Willemsen et al., 2018). This option is feasible if the robot's output is known beforehand, as is the case when manually adding gesture tags, so that the timing can be **manually annotated** as well (Figure 2.6). The timing of gestures can also be based on **heuristics** derived from human gesture studies, such as placing the starting time of the gesture 0.3 seconds before pronouncing the word it refers to (Aly & Tapus, 2016; Augello & Pilato, 2019). However, these heuristics do not consider differences between humans and robots, for example in the speed at which they can perform the gestures. To take the robot's physical abilities into account, there are implementations that pre-render the robot's utterance to an audio file via text-to-speech before actually playing this audio (Salem et al., 2013b; Shi et al., 2010). This file can be used to find out how long it will take the robot to pronounce certain words, which can be combined with knowledge of the robot's motor speeds to more accurately align the timing of co-speech gestures (Ng-Thow-Hing et al., 2010; Salem et al., 2013b; Shi et al., 2010; Figure 2.7). As with gesture selection, **contextual factors** can play a role in determining the optimal timing of a gesture. This includes information at the syllable level, for example by aligning the gesture stroke with the most accentuated syllable in a phrase (Mlakar et al., 2013), as well as handling scenarios where two planned gestures overlap in time, or in dialog settings where the robot's speech could be interrupted (Ishi et al., 2019).

### 2.4.4 Gesture synthesis

A recent trend in the planning of robot-performed gestures is to consider the design, selection, and timing steps as one single integrated step of synthesizing gestures from

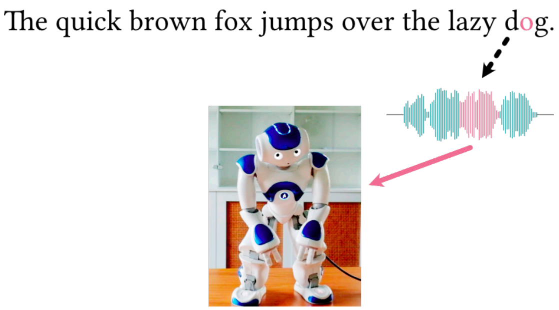


Figure 2.7: Gesture timing based on the robot's speech signal, combined with the robot's known motor speeds.

a trained mapping between an input signal — audio or text — to an output signal: the motions for the robot to perform. While one paper (Aly & Tapus, 2012) documents an implementation using coupled hidden Markov models (CHMM), the majority of the reviewed work follows scholars' increasing interest in (deep) neural network-based approaches. Sequence-to-sequence algorithms in particular are commonly used for the purpose of gesture synthesis. The **direct mapping from speech to motion** can be learned by training neural networks on the frames of pre-existing robot gestures (Marmpena et al., 2020; Marmpena et al., 2019; Rodriguez et al., 2019), motion capture data (Tuyen et al., 2020a, 2020b), but also on large, naturalistic corpora that were not collected specifically for the purpose of generating robot-performed gestures, including recordings of TED talks (Shimazu et al., 2018; Yoon et al., 2019; Yu & Tapus, 2020), or talk shows (Hua et al., 2019). As a pre-processing step, audio features such as Mel-frequency cepstral coefficients, log filter banks, pitch, or energy can be extracted from the speech signal (Aly & Tapus, 2012; Ondras et al., 2020; Shimazu et al., 2018). Alternatively, the speech signal can be transcribed to text (Hua et al., 2019; Tuyen et al., 2020a, 2020b; Yoon et al., 2019), which is then sometimes fed into a word embedding algorithm, such as word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b) or GloVe (Pennington et al., 2014). This enables the neural network to also generate motions that originally accompanied semantically similar, but not identical words to what it was trained on. At the same time, the pose of the people in the videos over time is extracted automatically using a pose tracking algorithm such as OpenPose (Cao et al., 2017). The neural network then learns a mapping from either auditory features or word vectors to gestures that can be performed by the robot (Figure 2.8). It is also possible to generate the robot's motion based on the interlocutor's motion, instead of the co-occurring speech signal, to read a social situation (e.g., a person is crying), and to have to robot reciprocate (e.g., by offering a hug) (Ko et al., 2020).

Because the gestures are learned from naturalistic examples, and neural networks are able to **generalize the existing motions to create new gestures** (Rodriguez et al., 2019), the resulting motions are more varied than those that are designed beforehand (Ondras et al., 2020; Yoon et al., 2019), and they are also generally perceived as more 'vivid' (Shimazu et al., 2018). Additionally, because the training process can be automated and the synthesis generalizes to new text input, this approach is less labor intensive than the gesture design and selection methods discussed previously. One of the main challenges with a neural network-based

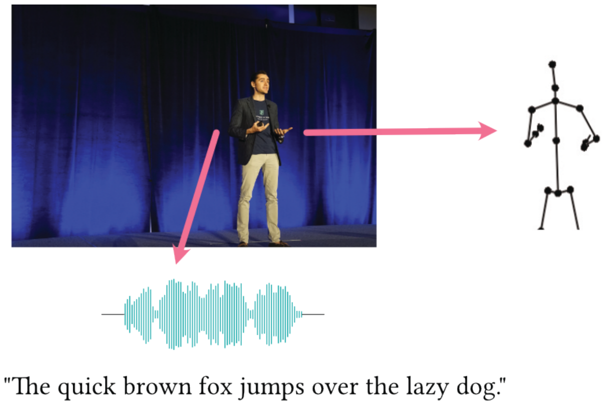


Figure 2.8: Gesture synthesis, where (transcribed) audio and pose data are automatically extracted from video recordings on a large scale, which can then be used for the robot's gesture production process.

approach however, is that the resulting gestures can be incongruent with the meaning that is conveyed through speech (Ondras et al., 2020), particularly when prosodic cues are used as input instead of the content of the utterance. This can be solved by taking a hybrid approach, where the synthesized gestures are augmented with several key frames from a gesture that was selected from a lexicon, based on word meaning (Shimazu et al., 2018). In addition, with neural networks there is less control over the types of gestures that are included, and it is more difficult to include contextual information, compared to the aforementioned rule-based, predictive, and probabilistic approaches.

### 2.4.5 Adjusting motion parameters

As humans, our gestures will never look identical: there is natural variation in our motions even if we perform the same shape twice, we tend to change our gesturing behavior based on whether our communication is successful or not (Hoetjes et al., 2015), or on established common ground (Mertens & Rohlfing, 2021). Furthermore, the way we gesture can be influenced by our personality and emotional state or mood (Dael et al., 2013; Hostetter & Potthoff, 2012; Kipp & Martin, 2009), as well as our age (Jain et al., 2016). Particularly for children, the type of information that is described and differences in cognitive abilities (Abramov et al., 2021) further affect their gesturing behavior. In addition, we take into account contextual information and use this to, for example, target our pointing gestures at the right referent (Haviland, 2000), or mirror the postures and behavior of our interlocutor to build rapport (Lakin & Chartrand, 2003). These factors that influence our gesture production process have been implemented and studied in the context of robot-performed gestures as well. For example, it is conceivable that a robot that adds some degree of **variation** to its gesturing behavior is perceived as more human-like, natural, and intuitive (Gielniak et al., 2011), and could sustain people's interest in the robot for a longer period of time (Marmpena et al., 2019). Variation can be added by manually designing different gestures for the same concept (Paplu et al., 2020; Chapter 6), by using neural networks to generate such variations automatically (González et al., 2019; Marmpena et al., 2020; Marmpena et al., 2019), by introducing (random, but constrained) noise to the motion (Gielniak et al., 2011), or by building complex gestures out of basic, simple motions with a probabilistic approach to add variation (Sunardi & Perkowski, 2020).

Social robots could also use their gestures to **convey a certain emotion** such as happiness, sadness, anger, fear, or high level affective states such as excitement or arousal. This can be done by creating gestures that focus solely on conveying an emotion, for example a cheering motion to depict happiness (Aly & Tapus, 2020; Jung et al., 2004; Paplu et al., 2020; Viergutz et al., 2014), which can then be triggered using one of the gesture selection approaches discussed in Section 2.4.2, for example whenever an emotional word, such as *sadness*, is encountered either in the robot's or the interlocutor's speech (Aly & Tapus, 2020). An emotional component can also be added to existing gestures by altering their velocity (speed) or amplitude (size)[4] (Claret et al., 2017; Le et al., 2011; A. Lim et al., 2011; Ng-Thow-Hing et al., 2010; Prajod & Hindriks, 2020; van de Perre et al., 2018; J. Xu et al., 2013). After using a neural network to synthesize gestures, researchers found that the resulting 'gesture space' contained examples of the same gesture with varying amplitude, therefore the emotional component of the gesture could be altered by sampling from different areas in this gesture space (Marmpena et al., 2020; Marmpena et al., 2019). Next to motion characteristics, the robot's pose (e.g., hand height, palm facing upward or downward, fingers stretched) has been shown to play a role in conveying different emotions (J. Xu et al., 2013). In one study, the robot used one arm to perform the original gesture, and the other to display an emotion, for example by pointing while covering its face to express fear (van de Perre et al., 2018). Both pose and motion can also be used to

[4] J. Xu et al. (2013) note that these modifications to existing, task-based gestures tend to convey long-term mood, while specifically designed gestures convey short-lived emotional states. However, the work that is discussed in this section appears to use the terms emotion and mood interchangeably.

convey a certain personality, such as introversion versus extroversion (H. Kim et al., 2008; Stolzenwald & Bremner, 2017), or dominance (Peters et al., 2019).

Part of a social robot's intelligence stems from its ability to **observe its physical and social surroundings**, and to optimize its gesturing behavior based on these observations. Therefore, a number of sensing capabilities have been implemented to improve the accuracy of a robot's pointing gestures, including object localization (Gulzar & Kyrki, 2015; Lemme et al., 2013), region identification (Hato et al., 2010), and social context recognition (e.g., presence and positioning of people) to ensure socially appropriate pointing (Ishi et al., 2020; Liu et al., 2017). Other gesture types, such as iconic gestures, may also need to be modified if the interlocutor is not directly facing the robot (Kondo et al., 2012; Tay & Veloso, 2012). The robot may also have to change its position or orientation before performing gestures, in order to target them correctly (Gulzar & Kyrki, 2015; Shi et al., 2010; Tay & Veloso, 2012). In addition to targeting their gestures, robots can monitor and adapt to the interlocutor's or user's emotional state (González et al., 2019; Jung et al., 2004; Tuyen et al., 2021; Valenti et al., 2020) or personality (Aly & Tapus, 2016; Stolzenwald & Bremner, 2017), their level of engagement with the interaction (Szafir & Mutlu, 2012), or auditory feedback received from an audience (Kraemer et al., 2016). As further discussed in Section 2.5.6, people with autism spectrum disorder (ASD) could benefit from training non-verbal communication skills with a robot by means of imitation. However, autistic people are known to vary in their imitation and motor coordination skills, which is why researchers have built a system that is able to measure a person's ability to mimic the robot's gestures, and to adjust parameters of the robot's movements (e.g., amplitude, velocity, and frequency) if needed (Ranatunga et al., 2015). Finally, in some cases the robot is also made 'self-aware', for example of any objects it may be holding that prevent it from gesturing (Holroyd et al., 2011), cost in terms of motor power required (Holroyd et al., 2011), joint failure (Jutharee & Maneewarn, 2016), or its current positioning (e.g., sat on a chair; Rodriguez et al., 2018).

### 2.4.6 Summary

In this section we have provided an overview of the state of the art in gesture planning. Similar to the design of robot-performed gestures (Section 2.3), we can distinguish between two different approaches: manually annotated, and automated using rule-based or artificial intelligence-based approaches. Gesture synthesis can be seen as the extreme case of an automated approach, as it integrates the design and planning of the gesture production process by inferring the motions themselves, their mapping to speech, and contextual information from large sets of data. In addition, these implementations can generalize to create new gestures, and they are able to perform gesture selection and timing for utterances they were not trained on. Following the philosophy behind artificial neural networks that power the majority of gesture synthesis implementations, these systems could be considered as most closely resembling the way we as humans (subconsciously) decide when and how to gesture.

The most suitable gesture planning approach depends on the use case: manual and rule-based planning provide more control over the robot's behavior than gesture synthesis, and thus result in a more predictable, constrained interaction, where contextual factors can be taken into account. This is desirable if the robot's gestures need to be accurate and relevant to the message that is conveyed through speech, for example in the case of using gestures to support a robot's teaching efforts. For general social conversation, the interaction might benefit from the added variation and scalability of automated approaches. In the next section, we will discuss experimental studies that have investigated the effects of robot-performed gestures, in which most studies rely on either manual or rule-based automated planning to ensure consistency between interactions.

## 2.5 Effects of robot-performed gestures

### 2.5.1 Methodological characteristics of experimental studies

Our review of existing literature resulted in **124 papers** that discussed the (potential) effects of robot-performed gestures based on a human-robot interaction study. From these 124 papers, we identified the following characteristics:

- **Focus/theme:** 26 papers discuss robot-performed gestures for communicative purposes (Section 2.5.2), 69 investigate the perception of the robot and interaction (Section 2.5.3), 18 examine engagement with the interaction (Section 2.5.4), 36 focus on task performance (Section 2.5.5), and 11 cover gestures for interactees with special needs (Section 2.5.6);

- **Gesture types produced by the robot:** 58 emblematic (symbolic, meaning is agreed upon), 50 deictic (pointing), 31 iconic, 16 metaphoric, 12 beat, 3 adaptors, 3 sign language (12 unknown);

- **Gesture theory references:** 75 out of 124 papers (60%) explicitly cite studies on human gesture production;

- **Commonly used robots:** NAO (68), Pepper (11), Robovie (5), Wakamaru (5), ASIMO (3), Erica (3), and Sota (3);

- **Gesture design approach:** 62 manual, 19 from demonstration (47 unknown);

- **The quality (i.e., comprehensibility) of the gestures' design** was assessed in 26 out of 124 papers (21%);

- **Gesture planning** (i.e., selection, co-speech timing, adjusting motion parameters) was done manually (e.g., Wizard of Oz, predefined scripts, or prerecorded videos) in 70 cases, and was automated (e.g., using rule-based systems or neural networks) in 39 papers (15 unknown);

- **Participants** in the studies were adults in 93 papers, children or teenagers in 28, elderly in 5, and interactees with special needs (e.g., learning problems) in 2 (7 unknown);

- **Sample sizes** ranged from 6 to 7685 ($M = 193, Mdn = 32, SD = 833$) (3 unknown);

- **Group size:** 112 were one-on-one sessions with the robot, 2 were with pairs of participants, and 7 were groups of at least 3 participants, or public spaces where any number of people could be drawn to the robot simultaneously (3 unknown);

- **Number of sessions:** 108 single session experiments, 16 consist of multiple sessions (ranging from 2–12 sessions, $M = 5, Mdn = 4, SD = 3$);

- **Location:** 75 papers describe lab studies, 31 field studies, and 14 online studies (5 unknown);

- **Effects found:** 70 papers report a positive effect of gestures, 30 neutral (e.g., not focusing on comparing gestures to no gestures) or no effect, 15 found mixed results, and 9 report a negative effect.

A table summarizing key details of the papers covered in this chapter is included in Appendix 2.C[5], and the full table is available in the supplementary materials[6].

### Measuring the effects of robot-performed gestures

The diverse effects of robot-performed gestures that are covered in this section were measured in multiple different ways, including both quantitative and qualitative measurements, and focusing on both attitudinal and behavioral aspects of a participant's interaction with the robot. Of the 124 papers covered in this part of the review, 53 (43%) present a between subjects study, 44 (35%) within subjects, 7 (6%) have a mixed design, and 19 (15%) were exploratory studies without experimental conditions (1 unknown). Furthermore, 27 (22%) used observations to study participants' behavior or emotional state, 10 (8%) used focus groups or interviews, 50 (40%) measured a form of task performance (e.g., accuracy, recall, response time), and 14 (11%) used other methods, such as automatic logging, content analysis, or implicit measures, such as electroencephalography (EEG), eye tracking, or automatic facial expression analysis. Finally, the majority of studies, 90 (73%), used self-report measures, generally in the form of a questionnaire. Of these 90, 66 (73%) constructed their own questions specifically for their study, while 37 (41%) used existing measurement scales, such as the Godspeed scale (Bartneck et al., 2009), Negative Attitudes Towards Robots Scale (NARS; Nomura et al., 2006), or Self-Assessment Manikin (SAM; Bradley & Lang, 1994) — 13 (14%) used a combination of both custom and existing scales. The majority of these self-report measures, and also of the evaluation methods in general, appears to result in quantitative data. Out of the 124 papers, 55 (44%) combined at least two of the aforementioned measures.

## 2.5.2 Gestures for communicative purposes

In this section, we look at communicative gestures: gestures that are produced intentionally to convey information, facilitate the flow of conversation, or to indicate the robot's intentions. Particularly for iconic, metaphoric, and emblematic gestures, it is important that their meaning is understood by the interlocutor for them to reach their communicative goals. For gestures that serve a goal in facilitating human-robot communication, such as the use of adaptors to help indicate the desire to end a

[5] Note: this table contains 125 references instead of 124: Because one article (Aly & Tapus, 2020) describes the same study as another (Aly & Tapus, 2015) it was only included once in the description of characteristics, but both are in the table.

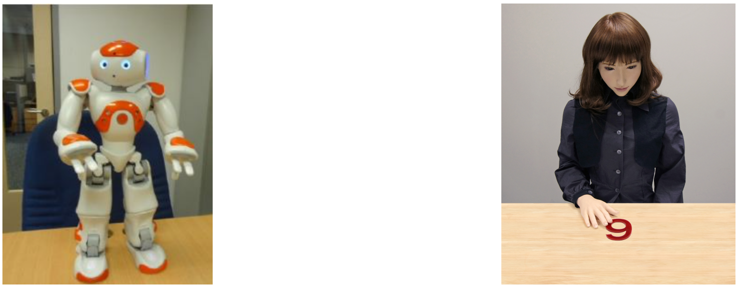
[6] https://osf.io/uj9fq/

conversation, it is important to investigate whether they are successful at reaching these communicative goals.

A number of studies have verified whether the information conveyed through robot-performed iconic, metaphoric, and emblematic gestures is understood by people, in spite of the robot's limited capability to perform these gestures in terms of smoothness and degrees of freedom. These studies generally present positive results, where **most robot-performed gestures are correctly understood by participants** (Bremner & Leonards, 2015a; Cabibihan et al., 2012; M. Zheng et al., 2019; M. Zheng & Meng, 2012). While in some studies the comprehensibility of robot-performed gestures was on par with human-performed versions (Bremner & Leonards, 2015a), most of the studies found that human-performed gestures are better understood than their robot-performed versions (Cabibihan et al., 2012; M. Zheng et al., 2019; M. Zheng & Meng, 2012). Cognitive processing of human-performed gestures has been shown to happen automatically (Kelly et al., 2010a), however a study has found that this does not appear to be the case for robot-performed gestures (Hayes et al., 2013). There does seem to be a relationship between people's aptitude at understanding human-performed and robot-performed gestures (Riek et al., 2010), and research further indicates that robot-performed gestures might become easier to understand once people's familiarity with robots increases (M. Zheng et al., 2019). An exploratory study reported that gestures described as 'simple' and 'relatable' were easier to comprehend than more complicated and abstract ones, that the actuator sounds from the robot's moving limbs were distracting, and that other modalities such as facial expressions could help clarify the robot's message (Abdul Jalil et al., 2012). For cooperative gestures (e.g., beckoning), abrupt and front-facing gestures evoked faster responses than smooth and side-oriented ones (Riek et al., 2010).

Although the ability for robots to accurately **refer to objects or locations** using deictic (pointing) gestures does not reach human-like performance (St. Clair et al., 2011), our review identified five factors that improve pointing accuracy: The target is easier to identify (1) if the robot is pointing at objects that are on the same side as the pointing arm, (2) when participants stand behind the robot instead of next to it, (3) if the robot also orients its head toward the target, (4) if the robot is able to actually touch the object, and (5) with practice over time (Bennewitz et al., 2005; Sauppé & Mutlu, 2014; St. Clair et al., 2011; X. Wang et al., 2014).

Given the fact that gestures are generally understood correctly, researchers have



(a) Using gestures to give feedback ('Please calm down') as a social mediator (from Tahir et al., 2020, image reprinted with permission from the original authors).

(b) Using a reaching gesture to support perspective taking (from Zhao and Malle, 2019, image reprinted with permission from the original authors).

Figure 2.9: Examples of studies regarding the use of gestures for communicative purposes.

started exploring **various communicative roles** that these gestures could fulfil. This includes using gestures to acknowledge a request for service (although gaze was found to be a stronger cue; Yamazaki et al., 2016), to trigger curiosity and topic exploration (Meena et al., 2012), to further clarify verbal feedback given as a social mediator (Tahir et al., 2020; Figure 2.9a), and to facilitate a natural way to end a conversation (Isaka et al., 2018). Beat gestures, similar to their role in human-human communication, can be used to emphasize parts of the robot's speech, although at a lower success rate than human-performed beat gestures (Bremner & Leonards, 2015b). Two studies have shown that a robot's use of reaching gestures can stimulate visual perspective taking (Zhao et al., 2016; Zhao & Malle, 2019). The degree to which perspective taking occurs appears to be linked to how human-like the robot is in appearance, where human-like robots such as Erica (Figure 2.9b) can stimulate visual perspective taking as well as a human can, while robots that have a less human-like appearance, such as NAO, can only do this to a lesser degree (Zhao & Malle, 2019).

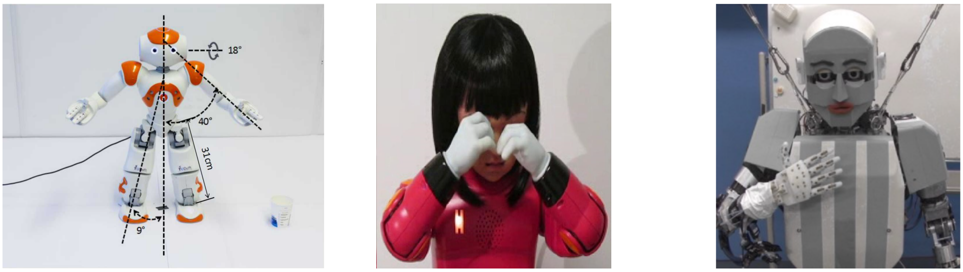
As in human-human communication, robots can mirror the interlocutor's gesture characteristics (e.g., speed) in order to build rapport (Stolzenwald & Bremner, 2017). This mirroring also happens the other way around, where it has been observed that human interlocutors mirror the timing (Robins et al., 2008), exact movement (Bao & Cuijpers, 2017), emblematic gestures (Nalin et al., 2012), or deictic gestures (especially when presented in combination with gaze; Bennewitz et al., 2005; Iio et al., 2011) performed by the robot. Another line of research investigated robots as proxies for people who are unable to gesture themselves, such as people with Parkinson's disease (Valenti et al., 2020), or tetraplegics (Kashii et al., 2016). In the latter case, it was found that interacting with the gesturing robot proxy mostly improved the experience for the hypothetical tetraplegic patient (no actual patients were recruited), and not the interlocutor. Finally, two robots can also gesture to each other, to let people in the environment know that they are able to communicate with the robots via gestures (Kanda et al., 2002). Interestingly, although gestures have been found to play an important role in turn taking behavior in human-human communication, this appears to be unexplored in human-robot interaction studies (Skantze, 2020).

### 2.5.3 Gestures' effects on the perception of the robot and the interaction

Gestures and the way they are performed can affect how we are perceived by others. For example, various aspects of our personality (Hostetter & Potthoff, 2012), such as extroversion, as well as our mood or emotional state (Dael et al., 2013; Kipp & Martin, 2009), are represented in how we gesture. We therefore expect that the perception of a robot and the interactions with this robot are also shaped by whether and how it uses gestures to communicate. This perception of the robot is most commonly measured using self-report, for example by asking participants to indicate the personality type or emotion they think the robot is expressing, or by using a questionnaire (e.g., Godspeed; Bartneck et al., 2009) to measure, among others, the likeability and human-likeness of the robot. In this section we will cover four different elements of the way the robot is perceived by others: personality, emotional state or mood, human-likeness, and overall likeability combined with enjoyment of interacting with the robot.

Researchers have successfully incorporated a number of **personality types** into the robot's gesturing behavior, including Big Five traits (Bremner et al., 2016), extroversion and introversion (Aly & Tapus, 2016; H. Kim et al., 2008; Ligthart et al., 2019), thinking versus feeling (H. Kim et al., 2008), and dominance (Li et al., 2015; Peters et al., 2019). Extroversion is usually accomplished by modulating the gestures' speed, size, and frequency (H. Kim et al., 2008; Ligthart et al., 2019), while dominance relates to the size of the gestures, or expansiveness of the robot's posture (Peters et al., 2019; Figure 2.10a). It was found that a less energetically gesturing, introverted robot elicited more self-disclosure from both introverted and extroverted children, compared to an extroverted robot (Ligthart et al., 2019). In another study, a robot that used larger and faster gestures was perceived as more extroverted, and was reported to be more enjoyable and capable (H. Kim et al., 2008). It was also shown that people might prefer a robot that matches their level of extroversion (Aly & Tapus, 2016). Together with its appearance and voice, gestures are found to be an important cue to recognize a robot's personality and individuality (Mikata et al., 2019).

It is also possible for robots to use their gesturing capabilities to **express various emotions**, including anger, happiness, fear, and sadness (Aly and Tapus, 2015, 2020; Augustine et al., 2020; Claret et al., 2017; Kaushik and Simmons, 2021; Prajod and Hindriks, 2020; Rehm et al., 2016; Tielman et al., 2014; Tsiourti et al., 2017; Figure 2.10b), as well as arousal (Claret et al., 2017; Marmpena et al., 2020; Prajod & Hindriks, 2020), valence (Marmpena et al., 2020; Prajod & Hindriks, 2020; Valenti et al., 2020), and mood (J. Xu et al., 2014, 2015a, 2015b). Literature has shown differences in how well certain emotions, conveyed through robot-performed gestures, can be recognized compared to others, where happiness and sadness generally tend to be easier to recognize compared to calm, anger, fear, and surprise (Claret et al., 2017; Kaushik & Simmons, 2021; Prajod & Hindriks, 2020; Tsiourti et al., 2017). How well humans are able to recognize emotions from the robot's gestures further appears to be dependent on demographic factors, such as previous experience with robots, cultural background, age, their level of introversion–extroversion, and gender (Aly & Tapus, 2015, 2020; Kaushik & Simmons, 2021; Rehm et al., 2016; Tuyen et al., 2021). For example, extroverted participants in one study rated the robot's emotional behavior



(a) The robot can change its posture and gesture characteristics to appear more dominant (from Peters et al., 2019, image reprinted with permission from the original authors).

(b) Gestures can be used to display certain emotions, in this case sadness (from Aly and Tapus, 2015, image reprinted with permission from the original authors).

(c) People tend to find robots that match their culture's gestures more likeable (from Trovato et al., 2013, image reprinted with permission from the original authors, copyright Takanishi Lab).

Figure 2.10: Examples of studies into how gesturing affects the way the robot is perceived.

as more expressive than introverted participants (Aly & Tapus, 2020). In addition, in this same study — which used a female (ALICE) robot — male participants generally found the robot's emotional behaviors more expressive than female participants. The authors postulate that this may be caused by the opposite-sex attraction of participants to robots (Aly & Tapus, 2020).

Emotion-expressing gestures can be implemented by designing specific gestures that portray a certain emotion (imitating crying to indicate sadness), or by modulating the speed and size of existing gestures (Marmpena et al., 2020; J. Xu et al., 2015a). By matching its mood, expressed through gestures, to the mood of a story it is telling, the robot is able to come across as more enthusiastic (J. Xu et al., 2015a). A study with gesture synthesis shows that if a mapping between prosodic speech information and gestures is extracted from TED talks, the resulting gestures performed by the robot are perceived as enthusiastic as well (Shimazu et al., 2018). In a study comparing between a humanoid robot (NAO) conveying emotion through gestures and a non-humanoid robot (Jibo) using animations on its screen to show emotions, children indicated that they prefer emotion conveyed through gestures from the humanoid robot (Émond et al., 2020). Furthermore, a robot that aligns its emotional gestures with a child's perceived emotional state was evaluated more positively than one that did not adapt to the child, and the adaptive robot in turn elicited more expressive behavior (e.g., smiles and frowns) from the child (Tielman et al., 2014).

Gestures generally **make the robot appear human-like or natural** (Bennewitz et al., 2005), especially compared to a robot that does not gesture (Asselborn et al., 2017; Carter et al., 2014; Hasegawa et al., 2010; Huang & Mutlu, 2014; Liles et al., 2017; Okuno et al., 2008; Ondáš et al., 2017; Park et al., 2011; Salem et al., 2013a; Tahir et al., 2020), although this depends on the design and implementation of the gestures (Ahn et al., 2013; Shimazu et al., 2018), and how well their timing and speed match the co-occurring speech (Ghosh et al., 2019). Gesturing robots tend to be seen as social agents, especially when they also use gaze (Bao & Cuijpers, 2017) or imitate the gestures of the interlocutor (Burns et al., 2018). Research further indicates that there may be a moderating effect of the perceived human-likeness (anthropomorphism) of a robot on the effects of perceived *competence* and *warmth* — two factors that play an important role in building trust with other humans as well as with robots (Christoforakos et al., 2021). A recent meta-analysis by Roesler and colleagues further supports this effect, as they showed that anthropomorphism generally results in a higher degree of trust toward, and acceptance of robots (Roesler et al., 2021).

While one study in our review found that a robot that used congruent gestures was perceived as more human-like than one that performed random movements (K. Xu, 2019), in another study a robot that used incongruent gestures was rated as more human-like than one that used congruent gestures (Salem et al., 2013a). Next to the effects of congruent and incongruent gestures, the degree to which gestures are exaggerated can be used to make the robot seem more machine-like, human-like, or cartoon-like (Gielniak & Thomaz, 2012). It is thus conceivable that gestures also play a role in navigating the 'Uncanny Valley', where a robot or character that is seen as extremely human-like (but not quite human) can be experienced as creepy or eerie (Mori, 1970), which is often studied based on appearance. Initial research indicates that the uncanny effect also occurs when a robot uses head nodding without gestures (Thepsoonthorn et al., 2021). The addition of gestures improved the NAO robot's human-likeness and affinity ratings, moving beyond the uncanny valley, particularly when these gestures were designed from demonstration (using Kinect for real-time motion tracking), rather than manually designed (Thepsoonthorn et al., 2021). Several studies found no differences on perceived human-likeness or naturalness between a robot that did or did not gesture (Aly & Tapus, 2015; A. Kim et al., 2012), between congruent and incongruent gestures (A. Kim et al., 2012; Vogt et al., 2017b), only deictic gestures compared to deictic and iconic gestures (van den Berghe, de Haas, et al., 2021), or between a fixed, more frequent, or varied number of repetitions of a motion (Seo et al., 2014, 2015). Pointing gestures were rated as more natural if the robot also used gaze (Iio et al., 2011). Furthermore, humans are able to tell whether a robot's pointing behavior is controlled by, or based on, human behavior (Wykowska et al., 2015), and additional facilitation processes (e.g., establishing joint attention) also lead to more natural pointing interactions (Sugiyama et al., 2007).

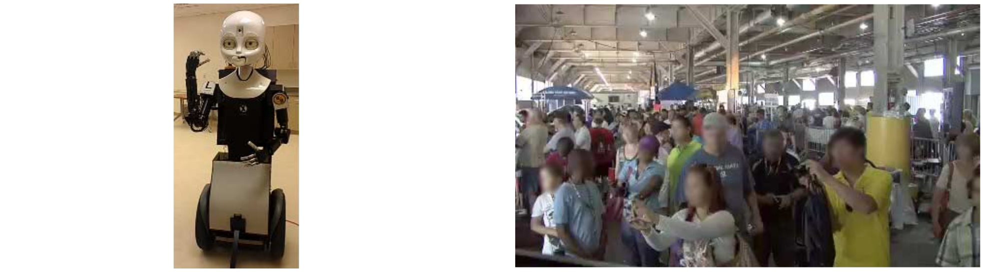
Finally, most studies that looked into the perception of the robot report positive effects of gestures on the robot being perceived as likeable, active, lively, enthusiastic, sympathetic, friendly, fun-loving, attractive, socially intelligent, credible, and people were more willing to accept the robot, and to keep using it in the future (Aly & Tapus, 2016; Asselborn et al., 2017; Huang & Mutlu, 2014; A. Kim et al., 2013; Liles et al., 2017; Meena et al., 2012; Moro et al., 2019; Ondáš et al., 2017; Salem et al., 2013a; Salem et al., 2012; Salem et al., 2011; Tahir et al., 2020; Tielman et al., 2014; J. Xu et al., 2015a; K. Xu, 2019), particularly when a robot's cultural gestures matched the cultural background of the interlocutor (Trovato et al., 2013; Figure 2.10c), when the robot did not act dominantly (Li et al., 2015), if it was using extroverted or exaggerated gestures (Hsieh et al., 2020; H. Kim et al., 2008), and when a motion (waving) was repeated more frequently (Seo et al., 2015), or a varying number of times (Seo et al., 2014). Participants in studies also reported a higher degree of shared reality (Lohse et al., 2014; Salem et al., 2013a), physical and social (tele)presence (Groechel et al., 2019; Kawaguchi et al., 2016), and being more empathetic with a robot that gestured (Burns et al., 2018; Sakamoto et al., 2005). Interactions with a gesturing robot were **rated more positively and found to be more enjoyable** (Carter et al., 2014; Pollmann et al., 2020), especially if the interlocutor was invited to mimic the gestures (Ligthart et al., 2020), or if the robot used listener-oriented gestures or aligned with the listener when giving directions (Hasegawa et al., 2010; Ono et al., 2001). For elderly, a robot that adjusted its gesture timing to their utterance speed was rated more positively than one that had fixed timings (Muto et al., 2009). Differences appear to exist between congruent and incongruent gestures, where a robot that used congruent gestures was perceived more positively (Goto et al., 2020; Wicke & Veale, 2020), and specifically as more sympathetic (Salem et al., 2012), while a robot that used incongruent gestures was perceived as more engaged and communicative compared to when gestures were absent (Salem et al., 2012). A robot that used incongruent gestures was found more likeable, and rated higher on intimacy and involvement than one that used congruent gestures (A. Kim et al., 2012; Salem et al., 2013a). One study showed no difference in likeability between a robot that used congruent or incongruent gestures (Vogt et al., 2017b). Researchers have also shown that the size and speed of the robot's gestures affect how the robot is perceived in terms of animacy, anthropomorphism, likeability, and the interlocutor's perceived safety (Deshmukh et al., 2018). Four studies found no effect of a robot's use of gestures on the perception of the robot (Admoni et al., 2016; Ham et al., 2015), specifically on trustworthiness or credibility (Huang & Mutlu, 2014; K. Xu, 2019). In one case, a robot that did not use gestures was deemed more trustworthy than one that did (Tielman et al., 2014). Another study indicates that a robot that uses exaggerated gestures, and gestures too frequently, could be perceived as confusing and irritating (Pollmann et al., 2020).

### 2.5.4 Gestures and engagement with the interaction

Engagement, often also referred to as *user* engagement in the field of human-computer interaction, can be defined as the level of involvement with an interaction:

how much attention and interest is invested by the user into this interaction (Lalmas et al., 2014; O'Brien & Toms, 2008). It can be subdivided into cognitive engagement (e.g., attention), affective engagement (e.g., emotional involvement), and behavioral engagement (or task engagement; e.g., complying with actions requested by the robot). There are several approaches to measuring these different types of engagement, including annotation, gaze tracking, or self-report. Robots could potentially use gestures to increase engagement from the interlocutor with the robot itself, as well as with the task at hand. As such, a robot that is more actively moving around will likely draw more attention than one that is static. Gestures can also support the robot's ability to express itself, potentially resulting in emotional involvement from the interlocutor. Finally, the robot's gestures may help to clearly communicate to the interlocutor about the tasks they are expected to perform, and to persuade them to actually perform these tasks.

A number of studies have looked at whether a robot's use of gestures supports its ability to **attract attention** from passersby in a public setting, where one study found no effect (K. Kim et al., 2017), one found a negative effect (Aizawa & Umemuro, 2021), while others found that people were more likely to engage with the robot, and listened to it for a longer time (Bremner et al., 2011; Moshkina et al., 2014; Figures 2.11a and 2.11b). These effects were only observed with congruent, human-like gestures and not with random movement (Bremner et al., 2011). Gestures have also been shown to help **maintain engagement during ongoing interactions** (Asselborn et al., 2017; Carter et al., 2014; Meena et al., 2012; Figure 2.12), particularly



(a) Gestures can help draw attention from a crowd (from Moshkina et al., 2014, image reprinted with permission from the original authors).

(b) A crowd is attracted by the (gesturing) robot from Figure 2.11a (from Moshkina et al., 2014, image reprinted with permission from the original authors).

Figure 2.11: Examples of studies into the effects of gestures on engagement with the robot and the interaction.

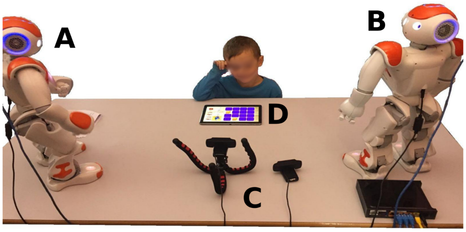


Figure 2.12: Adaptors while the robot is idle can help maintain engagement over time (from Asselborn et al., 2017, image reprinted with permission from the original authors).

in an educational setting (Ahmad et al., 2016b; De Carolis et al., 2019; Chapter 3; Chapter 6), and while a robot was assisting elderly with everyday tasks (Moro et al., 2019). This positive effect of gestures on engagement appears to persist over multiple sessions (A. Kim et al., 2013), and people indicated that gestures would likely retain their engagement on the long term (Wu et al., 2017). Exaggerated gestures in particular are perceived as more engaging than 'unexaggerated' gestures (Gielniak & Thomaz, 2012). No effect was observed regarding the presence or absence of actuator sounds on attention toward the robot (Jouaiti & Henaff, 2019). Several researchers have developed robots that can monitor the engagement level of an interlocutor or audience in real-time, and then use gestures to **regain attention** when this level drops (Bourguet et al., 2020a; Szafir & Mutlu, 2012). However, gestures can also draw too much attention: An eye tracking study showed that robot-performed gestures during a presentation draw more visual attention than similar gestures performed by a human presenter, and this was reported as a distracting factor, specifically due to the high frequency of gesturing and unnatural co-speech timing (Bourguet et al., 2020b).

### 2.5.5 How gestures can affect task performance

Humanoid, social robots are used in fields that involve frequent communication with humans, such as education, healthcare, and hospitality. We have seen that gestures can improve human-robot communication, help shape the way the robot is perceived, and increase user engagement. In addition, tasks that take place within these fields may also stand to benefit directly from a robot's use of gestures. For example, a robot's use of gestures can contribute to its success in **persuasion** (Chidambaram et al., 2012; Ham et al., 2015; Figure 2.13a), especially when combined with gaze (Ham

(a) Gestures can make the robot more persuasive, e.g., in a Desert Survival Task (from Chidambaram et al., 2012, image reprinted with permission from the original authors).

(b) Robots can also use gestures to support their teaching efforts (from Chapter 3, image reprinted with permission from the original authors).

Figure 2.13: Examples of studies regarding the influence of robot-performed gestures on task performance in various domains.

et al., 2015), or when polite gestures are used (N. Lee et al., 2017). This in turn can increase compliance to healthcare suggestions made by the robot (N. Lee et al., 2017), although another study did not find an effect on compliance (Moro et al., 2019).

Robots are also researched as tools in education, where their ability to use gestures is seen as an important feature by both teachers and primary school children (Ahmad et al., 2016a, 2016b), and a robot that uses gestures is perceived as a **better facilitator of learning** (Liles et al., 2017). It has been shown that gestures, particularly those that are congruent with what is conveyed via speech, can support second language tutoring (Vogt et al., 2017b; Vogt et al., 2019; Zhang and de Haas, 2020; Chapter 3; Chapter 6; Figure 2.13b), and the robot can also teach (culture-related) gestures (De Carolis et al., 2019). In one study, a robot used gestures to express either a positive or a negative mood. While this did not lead to a difference in quiz performance, the lecture with the positive robot received better ratings (J. Xu et al., 2014). In addition, researchers explored the idea of adding virtual (augmented reality) arms to a physical robot, but did not find improvements on an educational math task performance (Groechel et al., 2019). There may be differences in the effectiveness of gestures depending on learner characteristics: In a learning-by-teaching task, where the student had to correct mistakes in the robot's reading, robot-performed deictic gestures improved the ability of students with high reading proficiency to identify the robot's mistakes, but had a distracting effect on students with low reading proficiency (Yadollahi et al., 2018).

In the hospitality field, robots can use gestures to improve their ability to **give**

**directions or refer to objects or locations** (Ali and Williams, 2020; DePalma et al., 2021; Hasegawa et al., 2010; Lohse et al., 2014; Okuno et al., 2008; Ono et al., 2001; Figure 2.14), resulting in higher accuracy or speed of identifying the goal location, or better retention of the directions in the interlocutor compared to a robot that does not use gestures. There is however a potential adverse effect of robot-performed pointing gestures, as in one study participants also started pointing at the targets (in this case, other robots) instead of referring to them by name (Bennewitz et al., 2005). Also in the hospitality domain, a scheduling assistant robot that used gestures was rater higher on usability than one that did not use gestures (S. Lim et al., 2009).

Robot-performed gestures have further resulted in **better information retention** (Huang & Mutlu, 2013; van Dijk et al., 2013). This has been researched, for example, in the context of storytelling, where gestures can aid the memorization of story events (Gielniak & Thomaz, 2012; Szafir & Mutlu, 2012), particularly when these gestures are exaggerated (Gielniak & Thomaz, 2012). However, one study found that a robot's use of adaptor gestures while idle did not result in increased performance of children on a memory game (Asselborn et al., 2017). In another study, participants' retelling performance of an informative presentation did not improve if the robot used gestures (Huang & Mutlu, 2014). These mixed findings might be explained by several studies indicating that the benefits of gestures only apply to more difficult tasks (Admoni et al., 2016; Lohse et al., 2014), and in one case only when the robot was conveying a negative mood during the difficult task (J. Xu et al., 2015b). In addition, it is possible that gestures have to be congruent with what is communicated via speech for them to be effective: Several studies observed a negative effect of incongruent gestures on task performance (Huang & Mutlu, 2014;

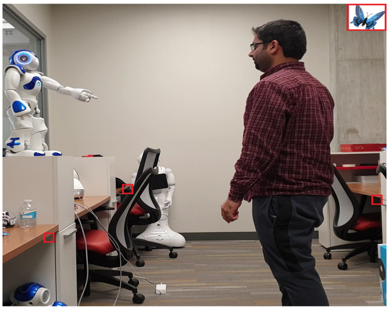
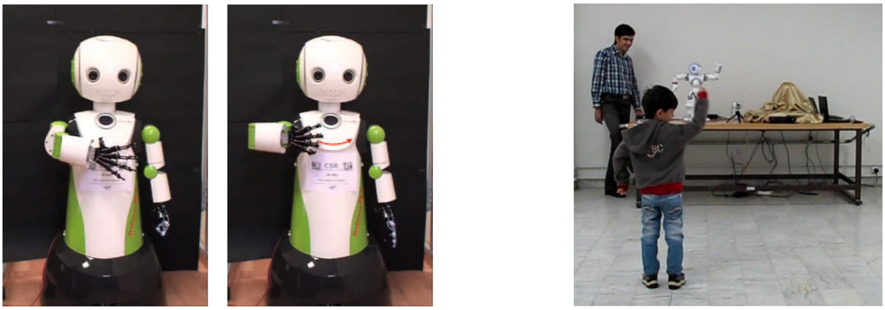


Figure 2.14: Particularly pointing gestures can help refer to objects or locations (from Ali and Williams, 2020, image reprinted with permission from the original authors).

Ono et al., 2001; Salem et al., 2013a), although two studies did not see a difference between congruent and incongruent gestures (Bremner et al., 2011; A. Kim et al., 2012). Finally, one study showed that the presence of actuator noise can negatively impact the performance on a rhythmic interaction (Jouaiti & Henaff, 2019).

### 2.5.6 Using gestures to support interactees with special needs

Our search uncovered one research group that is investigating whether social robots could be capable of communicating with humans using **sign language** (Akalin et al., 2013; Kose et al., 2015; Kose et al., 2012). This would enable people who are deaf or hard of hearing to interact with robots in a natural way. However, sign language remains a challenging endeavor as the currently available robots, including the NAO and Robovie R3 that are used in these studies, have limited degrees of freedom in their movement. While the NAO has only three fingers that can only move together in a gripping motion, Robovie has five fingers that can move individually, which is why it is generally preferred over the NAO for this purpose (Kose et al., 2015; Figure 2.15a). Initial research in this area shows that there are differences between various signs in how easy it is to recognize them (Akalin et al., 2013; Kose et al., 2012), and that children and teenagers find them more difficult to recognize than adults (Kose et al., 2015; Kose et al., 2012). Due to the robot's physical limitations, a number of signs risk ending up looking similar to others, making it more difficult to distinguish between them (Akalin et al., 2013).



(a) To a limited extent, robots can use (in this case Turkish) sign language (from Kose et al., 2015, image reprinted with permission from the original authors).

(b) Children with ASD can train gesture imitation with a humanoid robot (from Taheri et al., 2020, image reprinted with permission from the original authors).

Figure 2.15: Examples of studies into using robots to support interactees with special needs.

Social robots have also been researched as **therapeutic devices in the field of autism spectrum disorder (ASD)**, for example for training social skills or teaching gestures to nonverbal autistic people as an alternative means of communication. Robots show potential in this field, because they appear to be seen as something in between inanimate toys and fully human-like entities. As a result, they can provide a safe and relaxed environment for practicing social skills, while still ensuring a degree of realism in the interaction that allows the newly acquired social skills to carry over into human-human communication (Scassellati et al., 2012). Research shows that robots can be used to train gesture production and recognition with autistic children (So et al., 2019a; So et al., 2016; So et al., 2018a; So et al., 2019b; So et al., 2018b; Z. Zheng et al., 2016), and these skills are retained over time and can generalize to subsequent human-human communication (So et al., 2016; So et al., 2018a). Furthermore, gesture training with the robot in the context of storytelling resulted in improvements in the quality of the narratives children produced (So et al., 2019a). A robot and human teacher were found in two studies to be equally capable at training gesturing skills, but children were more likely to establish eye contact with the robot than with the human teacher (So et al., 2019b), and also tended to pay more attention to it (Z. Zheng et al., 2016). Other research however found that imitation performance among both typically developing and autistic children was better with another human, than with a robot (Taheri et al., 2020; Figure 2.15b). While there appear to be almost no differences between autistic and typically developing children at identifying the robot's emotions from facial expressions and gestures, fear in particular was found to be more difficult for autistic children to recognize (Salvador et al., 2015).

### 2.5.7 Summary

This section provides an overview of the different effects of robot-performed gestures on the interaction, the way the robot is perceived, and the goals that the robot tries to achieve. The results are generally positive: the meaning of robot-performed iconic, metaphoric and emblematic gestures is generally understood, and gestures can be used to fulfil a number of communicative roles that we also see in human-human communication, such as emphasizing parts of speech. Furthermore, robots can use their gesturing behavior to convey a certain personality, emotional state, or mood, and a robot that gestures is often perceived as more human-like than one that does not. Different motion parameters, such as the speed or size of the motions, can be adjusted to exercise some control over how the robot will likely be perceived: exaggerated motions could make the robot appear more cartoon-like and enthusiastic, and by increasing the size and speed of the gestures the robot will generally be perceived as more extroverted and dominant. Overall, a gesturing robot is also perceived more positively than one that does not gesture, people are more accepting of it and are more likely to engage with it over prolonged periods of time, and interactions are found to be more enjoyable. Finally, gestures can be used to support tasks in several different application domains, including healthcare, education, and hospitality. Robots and their gestures further show potential for supporting autistic people, particularly children, in training their (non-verbal) communicative and social skills, however it is as of yet unclear how these robots perform relative to a human trainer. Finally, another valuable application of social robots is in sign language, although the physical limitations of the currently available robots make it hard to distinguish between different signs.

Next to these overall positive and promising findings, there have been mixed results particularly relating to the use of exaggerated gestures, as well as congruency of the gestures with the information that is conveyed via other modalities, such as speech. There appear to be individual differences regarding a preference for exaggerated gestures, and this might also depend on the context, and the tasks that the robot is performing. Congruent gestures are generally preferred for improving task performance and often lead to positive ratings, but incongruent gestures might make the robot appear more likeable and human-like because of its perceived imperfections. Future work could further examine these nuances in the design of robot-performed gestures. In addition, we found little to no research exploring the effects of gestures in facilitating turn taking in conversations, and of the robot mirroring gestures or gesture properties (e.g., speed) from the interlocutor. Finally, there is no clear guideline on the desired overall frequency of the robot's gesturing, although some of the studies covered in this section seem to indicate that this frequency should be lower for robots than for humans, to avoid the movements being too distracting.

## 2.6 General discussion and conclusion

The aim of this literature review was to create a comprehensive overview of the state of the art in social robots that use hand gestures to support their communicative efforts. We focused specifically on (1) different implementations for the design and planning steps of a robot's gesture production process, and (2) the effects of robot-performed gestures on human-robot interaction in the various domains where social robots could be used. To ensure that the review is current and complete, we ran an exhaustive search, using broad search queries, on three relevant databases in the field of human-robot interaction. After screening the resulting papers and assessing full-text articles for eligibility, a total of 167 papers remained, of which 61 covered gesture planning approaches and 124 discussed an experimental study of the effects of robot-performed gestures. Information on the different approaches to gesture design was inferred from the set of 167 papers as well. In the sections below, we highlight the main results of the literature review, reflect on the state of the research field of robot-performed gestures, and introduce outstanding questions and avenues for future research based on these findings and reflections. The ten outstanding questions brought forward in this discussion have been summarized in Table 2.1.

### 2.6.1 Design of robot-performed gestures

The different gesture design approaches discussed in literature can be divided into two categories: manual design, and design from demonstration. **Manual design** is often done using key framing techniques known from character animation, where salient poses from a gesture are defined, and the robot smoothly moves between these poses when performing the gestures. In **design from demonstration**, motions are recorded from human performers and are then transferred onto the robot.

The main advantage of manually designing the robot's gestures is that the designer has full control over what the gestures will look like. This means that the robot's morphology and physical limitations can be taken into account, and instead of basing the robot's gestures on human-performed examples, the designers could also draw inspiration from animation theory, to make the gestures more exaggerated or cartoon-like. It may also be easier to modify manually designed gestures so that they convey a certain emotional state or mood. This increased level of control is likely the reason for the majority of the planning approaches and studies covered in this literature review to use manually designed gestures. Gestures that are designed from demonstration, however, are less labor intensive because they can be generated automatically on a large scale, and they tend to look more human-like because the entire motion, including its imperfections, is recorded from a human model and then performed by the robot. With portable depth sensors such as Microsoft's Kinect, and recent developments enabling pose tracking from arbitrary two-dimensional recordings, such as TED talks, we anticipate a move toward more gestures being designed from demonstration. At the same time, it would be interesting to explore whether **hybrid approaches** could succeed in combining the benefits of both gesture design methods, while avoiding their pitfalls (e.g., as initially explored in Shimazu et al., 2018). This leads to our first outstanding question in the field: *Will developments in sensor technology and AI improve the quality of robot-performed gestures collected from naturalistic data, and thus remove the need for manually designed gestures?*

The studies discussed in Section 2.5 highlight that the way in which gestures are designed plays a crucial role in their effectiveness, in terms of the way the robot is perceived, levels of engagement with the interaction, and the benefits to task performance. For example, exaggerated gestures can make the robot seem more cartoon-like, which in some studies resulted in a more positive response toward the robot (e.g., Hsieh et al., 2020; H. Kim et al., 2008), while in another study this led to confusion and irritation (Pollmann et al., 2020). It is therefore important that gestures are designed with a certain goal in mind (e.g., human-likeness, extroversion), that subsequent design decisions are carefully weighed, and that there is an **iterative design process** that evaluates whether the gestures succeed at reaching their goal before they are used. This holds for iconic, metaphoric, and emblematic gestures in particular: Since they are intended to convey a certain meaning, they need to be designed in such a way that this meaning is clearly understandable. Different strategies exist for performing iconic and metaphoric gestures, e.g., outlining the shape of a concept, or performing a certain action that is related to the concept (Müller, 2014). Which of these strategies is adopted may vary depending on the type of concept that is depicted, but also on characteristics of the person performing the gesture, such as their age (Jain et al., 2016). As a result, if the goal is for the robot to interact with children, it might be more effective to incorporate gestures that align with children's preferred gesturing strategies, and to also evaluate the quality of the gestures with children. However, a relatively small number of studies included in this review (21%) report on an evaluation of the comprehensibility of the gestures, before using them in their study. Given the substantial number of factors to consider at the design stage of a robot's gesture production process, combined with the lack of evaluations of this design step, we raise the second outstanding question: *How can we add structure and consistency to the process of designing and evaluating robot-performed gestures?*

Because most studies use the NAO or Pepper robots, and only six studies compare between different robots, it is as of yet unclear **how a robot's gesturing behavior relates to its physical appearance**. For example, machine-like gestures might be perceived as more natural on a mechanical-looking robot than a NAO robot, and it might appear unsettling for a robot that is human-like in appearance (e.g., Erica) to perform exaggerated, cartoon-like motions. Appendix 2.A contains images of all the robots that were used in the papers included in the literature review, and this shows substantial differences in physical appearance, in terms of how mechanical or human-like the robots look, and also in terms of their gesturing capabilities. It is conceivable that the robot's appearance, together with its use of gestures, determines how it is perceived and how the communication with others is shaped, although it is as of yet unclear how these two factors relate to and influence each other. According to one study, not only the way the robot as a whole, but also specifically its gestures are perceived may be affected by its appearance, as male participants rated a female robot's gestures as more expressive than female participants, which the authors attribute to opposite-sex attraction — a phenomenon from human-human interaction that appears to apply to human-robot interaction as well (Aly & Tapus, 2020). Because only limited research has investigated the relationship between a robot's physical appearance and gesturing behavior, the third outstanding question is: *How does a robot's physical appearance relate to its use of gestures, and how does the interplay of appearance and gesture influence human-robot interactions?*

### 2.6.2 Gesture planning

We identified four different themes within the planning step of a robot's gesture production process: gesture selection, co-speech timing, gesture synthesis, and adjusting the motions' parameters before executing a gesture, for example to tailor the gesture to a particular context. **Gesture synthesis** uses artificial intelligence to infer the design, selection, and co-speech timing of the robot's gestures from large sets of data. This is a promising avenue for future research, as it is now possible to use readily available sources of data, such as YouTube videos, as input, and these algorithms are able to generalize to generate new gestures, and to add gestures to new speech input. Interestingly, **co-speech timing** was only addressed by six papers, even though this is not a trivial task in complex real-world situations, such as free-form dialog, where the robot's speech can be interrupted. Incorrect timing can also lead the robot to be perceived as less human-like or natural (Ghosh et al., 2019).

Within the different planning approaches, we observe a similar distinction between **manual and automatic approaches** to gesture selection and timing as previously discussed for gesture design. Manual selection and timing can be done for example by annotating the robot's speech output to indicate exactly when a particular gesture is to be performed. Automatic, rule-based or artificial intelligence-based systems, including gesture synthesis, are less labor intensive because they map particular words (and, optionally, similar words) to gestures, and can even infer these mappings from datasets. In this case, too, we observe that the majority of experimental studies covered in Section 2.5 use manually planned gestures, for example by creating scripts for the robot to follow, by triggering gestures using a Wizard of Oz set-up, or by making video recordings of an interaction or even isolated gestures and then showing these to participants. This predominance of manual planning approaches is likely related to the large number of studies that took place either in the lab (61%) or online (11%), and the fact that we only included studies where the effects of the robot's gestures were investigated, where consistency between participants is desired in order to reliably study these effects. It is also easier with manual or rule-based approaches to take contextual information, such as the emotional state of the robot or interlocutor, into account.

Adjusting the motions' parameters before producing the gestures was the largest theme, with 36 papers investigating ways to add variation to the robot's gestures, or to tailor the gestures to a particular situation or context. A robot that is more contextually aware, for example by adjusting to the interlocutor's emotional state or by re-positioning itself to perform clearer pointing gestures, will appear to be more (socially) intelligent. The literature shows that it is possible to adjust gestures in order to express a certain mood, or to give the robot a particular personality. This could be tailored to the desired role of the robot, for example by making it appear more dominant to establish authority. Therefore, as with the design of the gestures themselves, **thought should go into the decisions involving gesture planning**, as these will also partly determine how the robot will be perceived and whether it will be accepted by the people interacting with it. For example, frequent gesturing makes the robot appear more extroverted, but if it gestures too frequently this can cause irritation or distraction (Bourguet et al., 2020b; Pollmann et al., 2020). Particularly **variation** in the robot's gesturing behavior is an interesting topic of studies, which has received some attention from a gesture planning perspective (e.g., Gielniak et al., 2011; Marmpena et al., 2019), but the effects of which have not been studied as elaborately yet (see, e.g., Chapter 6, for first explorations).

Compared to the design of the gestures, evaluations of the planning approaches are more common (78% of the papers). However, only 59% conducted an evaluation with human participants, and approximately half of these examined the effects of the planning approach on one or more aspects of the interaction between a human and the robot, and were therefore also discussed in Section 2.5 of this literature review. We believe it would be valuable to study the effects of these planning approaches *in situ*, but limited research appears to have done so. Therefore the fourth outstanding question is: *How do design decisions regarding gesture planning (e.g., introducing variation) affect the resulting overall human-robot interaction?*

### 2.6.3 Effects of robot-performed gestures on human-robot interaction

We presented a comprehensive overview of the various effects that robot-performed gestures can have on different aspects of human-robot interactions. These effects were divided into five different themes: communicative purposes, perception of the robot and the interaction, engagement, task performance, and effects specifically for interactees with special needs. The majority of papers discussed present **positive effects**, indicating for example that a robot that uses gestures is generally perceived as more likeable or enthusiastic (e.g., Salem et al., 2011), and that the robot's gestures can support tasks in a number of domains, such as education (e.g., Chapter 3). In addition, a number of neutral effects were found, for example when the perception of the robot *changed* because of its gestures, not necessarily for better or for worse (e.g., H. Kim et al., 2008). The predominance of positive effects could, however, be the result of publication bias (Rothstein et al., 2006), and there is a need for more replication studies in the field of human-robot interaction (Hoffman & Zhao, 2020; Irfan et al., 2018).

From studies into the **comprehensibility of robot-performed gestures**, we observe that they are generally understood by others, although not always as well as human-performed versions of the same gestures (Cabibihan et al., 2012; M. Zheng et al., 2019; M. Zheng & Meng, 2012). People's ability to understand the robot's gestures has been shown to depend on a number of factors, both related to the design of the gestures themselves (as discussed previously in Section 2.6.1), as well as individual differences in the observer. For example, a robot's gestures appear to be easier to understand for people that are more adept at interpreting gestures performed by humans (Riek et al., 2010), and familiarity with robots appears to play a role as well (M. Zheng et al., 2019). Preliminary research further indicates that the age of the observer may affect how well gestures are understood, where younger children tend to perform worse than older children, and both children and elderly worse than adults (Kose et al., 2012; Rehm et al., 2016; Chapter 6). Because the comprehensibility of the robot's gestures is likely a key prerequisite for them to be effective, particularly regarding task performance, we formulate the following fifth outstanding question: *Can we define the quality (e.g., comprehensibility, 'communicative success') of a robot-performed gesture, and can this be measured in a standardized, perhaps even automatic way?* A similar call for standardized evaluation metrics has been made specifically in the field of learning from demonstration (Argall et al., 2009), and recently quantified, computational evaluation methods (Zabala et al., 2021) as well as guidelines on how to conduct and report on gesture generation and evaluation for embodied conversational agents (Wolfert et al., 2021) have been proposed, as first steps toward addressing this outstanding question.

Integration of the robot's gestures with **other non-verbal modalities**, such as eye gaze or facial expressions, has been shown to aid the comprehension of the gestures (e.g., Abdul Jalil et al., 2012), and can lead to stronger beneficial effects compared to using gestures alone (e.g., Ham et al., 2015; Iio et al., 2011). In this literature review, we only included studies in which the effects of gestures were studied independently, so that we could assess the nature of these effects, and identify the role that (the design of) the robot's gestures played in the interaction. While it might be valuable to study the different modalities in isolation in the context of scientific research, in practice these could together be considered as the robot's socially intelligent behavior (Fong et al., 2003). Therefore, the sixth outstanding question is: *Should gesture generation be considered a separate system, or should we consider non-verbal multimodal output generation as one task?*

Overall, a robot's use of gestures has been shown to have a number of beneficial effects: the gestures can fulfill communicative roles (e.g., ending a conversation in a natural way), help refer to objects or locations, and allow the robot to convey a certain personality or emotional state. Furthermore, gestures can result in more positive ratings of the robot and the interaction, lead to higher levels of engagement, and are able to improve the robot's persuasive abilities, its teaching efforts, and information that is conveyed by the robot in combination with gestures is often better retained by an audience. From the literature covered in this review, we have identified a number of **factors regarding the design of the robot's gestures** that may affect how the robot is perceived by others, and how effective the gestures are at supporting various tasks. These factors include the previously discussed gesture properties such as speed or size, as well as the frequency of gesturing, and the extent to which the gestures are exaggerated or not. Another important factor is whether the gestures are congruent with the information that is conveyed via other modalities (e.g., speech), where congruent gestures appear to generally be preferred. Incongruent gestures can be used to make the robot appear more human-like and likeable, but this can come at the cost of reduced task performance. Although a number of these factors related to the design and implementation of the robot's gestures have been identified and studied, there is a lack of studies into how **individual differences in the observers** of the robot's gestures influence the effects that these gestures may have. Most studies are conducted with adults, from a single cultural background. There is reason to believe that there may be differences, particularly regarding the perception of the robot and its gestures, based on gender (Aly & Tapus, 2020) or cultural background (Trovato et al., 2013; Tuyen et al., 2021), which leads to the seventh outstanding question: *How do individual differences, e.g., based on cultural background, influence the effects of robot-performed gestures?*

An interesting theme that emerged from this review is **gesture mirroring**, both performed by the robot as well as by the interlocutor. In human-human communication, behavioral mimicry can be seen as an indication of rapport if it occurs subconsciously — known as the *chameleon effect* (Chartrand & Bargh, 1999) — and some have argued that this can be consciously used to increase rapport and liking (Chartrand & Bargh, 1999; Lakin & Chartrand, 2003). The fact that participants in several of the studies covered in this literature review started mirroring the robot's gestures (Bao & Cuijpers, 2017; Bennewitz et al., 2005; Iio et al., 2011; Nalin et al., 2012; Robins et al., 2008) could be seen as an indication that the robot is regarded as a social agent by the participants, with which they can form a social bond. This, in turn, could lead to long-term acceptance and engagement with the robot, which is important to establish lasting effects, for example in healthcare or education (Leite et al., 2013). One study illustrates that it is possible to have a robot mirror elements of the participants' gesturing behavior (Stolzenwald & Bremner, 2017), which could potentially be used as a tool to build rapport. Specifically in the field of education, having the student mirror the teacher's gestures can lead to increased learning outcomes compared to merely observing these gestures (e.g., Cook et al., 2008; Tellier, 2005), although to our knowledge this has not been researched with a robot as teacher. In short, limited research on gesture mirroring with robots shows promising results, and it would be worthwhile to investigate whether certain design decisions regarding the robot's gestures can stimulate mirroring by others. Therefore, we pose the eighth outstanding question: *Is mirroring of gestures, by the robot or by the interlocutor, always beneficial, and can we design the robot's gestures in such a way that they elicit more frequent mirroring from others?*

A further observation is that gestures generally make the robot appear more **human-like**. This is often considered a desirable quality, because it enables us to interact with robots in a way that is familiar and natural to us, and this tends to result in greater acceptance of these robots in our lives (Fink, 2012; Roesler et al., 2021). A more human-like robot, albeit in appearance in this particular study, was indeed shown to elicit more social behavior from humans in the form of visual perspective taking (Zhao & Malle, 2019), and human-likeness is mentioned as one of the components for stimulating child-robot relationship formation (van Straten et al., 2020). The meta-analysis by Roesler et al., 2021 also highlights the positive effects of human-likeness on likeability, perceived intelligence, activation, pleasure, trust, acceptance, task performance, and social behavior by the interactee, but mostly in the social domain — only some of these effects were found for the service and industrial domains. It is worth noting that this meta-analysis includes human-like features other than the robot's (gesturing) behavior, such as its appearance. It is as of yet unclear how the uncanny valley effect, where a robot that becomes too human-like can be perceived as eerie, applies to a robot's gesturing behavior. One study did investigate this effect, and found that human-like gestures can help move beyond the uncanny valley, but this was only tested with one robot platform (Thepsoonthorn et al., 2021). It is unclear whether human-likeness is always required and desired, or whether it is perhaps limiting the robots' potential to go beyond human abilities, and make use of their 'superpowers', such as having endless patience (Dörrenbächer et al., 2020). Furthermore, the fact that people tend to build social bonds with human-like robots could have a drawback of getting (too) emotionally attached to these robots (Zhao & Malle, 2019), and there are several ethical and societal implications to consider (Darling, 2017; de Graaf, 2016), particularly with regard to child-robot interactions (van Straten et al., 2020). The robot's gestures could, together with its appearance, potentially be used to navigate the scale of human-likeness, for example by making the gestures more exaggerated to position the robot as more cartoon-like, but further research is needed to verify to what extent this is possible, and desirable. This is the ninth outstanding question: *Under what circumstances should we aim for a human-like appearance and gestures, or rather consider robots a distinct entity with* *their own mode of expressing themselves (e.g., using exaggerated, cartoon-like motion)?*

Finally, robot-performed gestures show potential in **supporting interactees with special needs**. In the current literature review, we found promising results in the field of using robots as therapeutic devices for autism spectrum disorder (ASD), specifically to teach gestures to nonverbal autistic people and for training various social skills (see, e.g., Scassellati et al., 2012, for a review). Additionally, researchers have begun exploring whether the robot can communicate using sign language, to make robots more accessible to people that cannot use speech to communicate, such as the deaf or hard of hearing. However, particularly for sign language the currently available robot platforms appear to be limited in their degrees of freedom, which means that signs will have to be adjusted for the robot to be able to perform them, or they will be less clear compared to when they are performed by humans. Further research is needed to verify whether currently available robots are indeed able to communicate fully via sign language or, if not, to explore alternative options to make robots more accessible. One potential solution could be to use mixed reality, where virtual arms and hands can be used to allow for greater fluidity and freedom in movement compared to physical versions. We found one study that explored this option, but in the context of education and not sign language (Groechel et al., 2019). The need for accessibility in interactions with robots results in the tenth and final outstanding question: *How can we make robots more universally accessible: can we enable people with special needs to communicate with them in a meaningful and natural way?*

### 2.6.4 The state of the research field

Next to investigating the state of the art in robot-performed gestures and identifying outstanding questions, we have taken a critical reflective look at the way research in this field is conducted. A **positive trend** is that the majority of research covered in this literature review includes a controlled experimental study in which (aspects of) the robot's gesturing behavior was manipulated. This allows us to systematically identify the effects of gestures on human-robot interactions, and provides further qualitative information regarding the design decisions that may have influenced their effectiveness. Furthermore, 44% of the papers combine multiple measurement instruments to obtain a comprehensive overview of the effects of robot-performed gestures. Four methodological aspects that could be improved in the majority of the research covered in this literature review relate to replicability, external validity, the measurement instruments used, and connections with other disciplines. Based on our reflections, we propose four points of improvement addressing these methodological aspects, most of which have previously been suggested for the human-robot interaction field in general (Bethel & Murphy, 2010; Hoffman & Zhao, 2020).

Table 2.1: Outstanding questions in the study of robot-performed gestures.

1. Will developments in sensor technology and AI improve the quality of robot-performed gestures collected from naturalistic data, and thus remove the need for manually designed gestures?
2. How can we add structure and consistency to the process of designing and evaluating robot-performed gestures?
3. How does a robot's physical appearance relate to its use of gestures, and how does the interplay of appearance and gesture influence human-robot interactions?
4. How do design decisions regarding gesture planning (e.g., introducing variation) affect the resulting overall human-robot interaction?
5. Can we define the quality (e.g., comprehensibility, 'communicative success') of a robot-performed gesture, and can this be measured in a standardized, perhaps even automatic way?
6. Should gesture generation be considered a separate system, or should we consider non-verbal multimodal output generation as one task?
7. How do individual differences, e.g., based on cultural background, influence the effects of robot-performed gestures?
8. Is mirroring of gestures, by the robot or by the interlocutor, always beneficial, and can we design the robot's gestures in such a way that they elicit more frequent mirroring from others?
9. Under what circumstances should we aim for a human-like appearance and gestures, or rather consider robots a distinct entity with their own mode of expressing themselves (e.g., using exaggerated, cartoon-like motion)?
10. How can we make robots more universally accessible: can we enable people with special needs to communicate with them in a meaningful and natural way?

Regarding **replicability** of studies in the field of robot-performed gestures, we observe that crucial information is often missing, which prevents replication of the majority of the studies that are covered in this literature review. This mainly relates to a lack of clear descriptions regarding the design and (pilot) evaluation of the robot's gestures. It is often not clarified who designed the robot's gestures, and if they were based on human-performed examples, theoretical knowledge or best practices on how these should be designed for a particular target group. It is also not commonly measured how suitable the gestures themselves are at reaching their intended goal, before integrating them in an interaction. Furthermore, the process of integrating the gestures (i.e., planning approaches, frequency of gesturing) is often not documented in detail. Related to this point, the data, gestures (ideally in a platform-agnostic format), or other stimuli and measurement instruments are generally not made publicly available for others to use. We strongly believe that the research field would benefit from a move toward an open science approach.

The second point relates to the **external validity** of the studies that are covered in this review, where it is unclear whether the findings can be generalized to other robot platforms, contexts, or populations. We have identified four potential threats to external validity. First, studies use a number of different robots, that tend to differ greatly in terms of their physical appearance and gesturing capabilities (see Appendix 2.A for an overview). Particularly the NAO and, to a lesser extent, Pepper robots are frequently used, while the other robots only come up in very few studies. There is only a limited number of studies that compare between different robots, which is understandable given the high cost of these robots, as well as having to implement the same gestures across multiple platforms. However, this could hamper external validity as it is hard to gauge to what extent a particular robot's appearance and capabilities play a role in the findings from these studies, and how much is actually related to the gestures themselves. This could perhaps be solved by addressing our previous point, by making studies more replicable so that other researchers can perform them with different robots and across different contexts (e.g., cultures). Second, the majority of the studies consist of a single session, and they take place either in the lab, or online with prerecorded videos of the robot's behaviors. This begs the question whether the effects we observe in these studies persist over time. In addition, it is likely that the robot's gestures have a stronger effect if the robot is physically present, as opposed to using prerecorded videos (Li, 2015). Third, most of the studies involve one-on-one interactions, although we expect it is likely for robots to be approached by multiple people at once when they are used in daily life. Finally, oftentimes a convenience sample is used. Especially for research on robot-performed gestures to support people with special needs — which shows promising results, but is still relatively scarce — it is important to involve the right target group early on, and frequently.

Thirdly, a broad range of different **measurement instruments** is used to study the research questions set out in the papers. This cannot be avoided, because robots are studied in several different domains, and there are multiple aspects and outcomes related to people's interactions with robots (e.g., learning outcomes, perception of the robot). However, oftentimes self-constructed questionnaires are used to obtain self-report data. This, combined with the relative complexity and 'fuzziness' of the concepts that are measured, such as likeability and human-likeness, provides further challenges when trying to integrate the findings of multiple studies. By extension, it is challenging to study how these different concepts may relate to or affect each other, for example whether a more human-like robot is automatically also perceived as more likeable, or how people might show more (affective) engagement with a robot that is considered human-like and likeable. Also in this case, publicly sharing the measurement instruments used and raw data, as well as the use of validated questionnaires, would help address these challenges.

Finally, although a substantial number of the articles discussed (approximately 60%) does cite one or more seminal works from gesture studies, this is commonly done merely to introduce gestures and gesture taxonomies, and not to provide theoretical foundation for the study's design. We believe that it would be valuable to **connect more deeply with related fields of research**, such as human gesture studies, in order to further structure the research into robot-performed gestures. This will also allow for a more elaborate comparison between gestures performed by a robot and those performed by a human, e.g., in terms of cognitive processing of these gestures (Hayes et al., 2013), and whether we establish the same 'shared intentionality' with robots as we do with other people (e.g., Dennett, 1987; Tomasello & Carpenter, 2007). Future studies in this direction will also provide further insight into how robots are generally perceived by others — e.g., as an entity that is close to a human and perceived as a social agent (Bao & Cuijpers, 2017; Burns et al., 2018), or more as an inanimate object — and how gestures might potentially be used to influence this perception.

### 2.6.5 Conclusion

Hand gestures have emerged as a defining property of social robots and their physical embodiment and presence. By means of a systematic literature review, including 167 articles that met the inclusion criteria, we have created an overview of the state of the art regarding (1) the design and planning steps of a robot's gesture production process, and (2) the effects of the robot's gestures on the resulting human-robot interactions. Within the robot's gesture production process, there are manual as well as automatic approaches to the **design and planning** (e.g., gesture selection, co-speech timing) steps. Furthermore, in the planning stage it is possible to adjust several aspects of the gestures, such as their speed, to add variation or to adapt based on various types of contextual information (e.g., the robot's or interlocutor's emotional state).

Studies into the **effects of a robot's use of hand gestures** were divided into several themes: communicative purposes (e.g., visual perspective taking), perception of the robot (e.g., human-likeness, likeability), engagement, task performance, and supporting interactees with special needs. Articles across these different themes mostly present positive or neutral results, where studies with neutral results either did not compare with a robot that does not gesture, or found no difference between robots that do and do not use hand gestures. We can therefore conclude that it is important to incorporate gestures when designing a robot's **socially intelligent behavior**, as this will generally have a positive effect on the resulting human-robot interaction. This conclusion aligns with the essential role of gestures in communication between people (e.g., Clark, 1996). While gestures themselves are considered a crucial part of the robot's socially intelligent behavior (Fong et al., 2003), this literature review further shows that they can be used to facilitate other aspects of socially intelligent behavior as well, such as the robot's ability to express emotion and to build and maintain social relationships. In addition, we observed that the two topics of this review intertwine: **design decisions made while implementing the gesture production process appear to influence the effectiveness of the gestures**.

From the existing body of literature, we extracted **ten outstanding questions** that we believe could serve as a guideline for future work in the field. This includes monitoring developments in sensor technology and AI, adding structure to the gesture design and evaluation process, studying the relationship between a robot's physical appearance and gestures, investigating the effects of planning on the overall interaction, standardizing measurements of gesture 'quality', potentially integrating gestures with other modalities, incorporating individual differences, further studying gesture mirroring, discussing whether human-likeness is desirable, and ensuring universal accessibility of robots. Finally, in a **critical look at the research field** we observe that there are many well-designed studies that focus specifically on robot-performed gestures, and combine the use of multiple measurements to create a more comprehensive image of the effects of the robot's gestures. Based on this critical look, we propose four methodological points of improvement, which relate to replicability, external validity, measurement instruments used, and the need for connections with other disciplines. With these outstanding questions and suggestions, we aim to provide a concrete starting point for future research in this field.

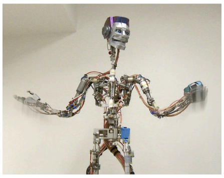
* * *

*In the current chapter, we provided an overview of the state of the art in robot-performed gestures. We observed that there are different ways to approach the design and planning steps of a robot's gesture production process, and that there are several potential effects of gestures on human-robot interactions, such as increased levels of engagement, or better performance on joint tasks. Not much research has been done in the field of education, and the effects of a robot's gestures on second language learning in particular have remained largely unexplored. In the next chapters, we therefore present several studies that do focus on this field, where we measured how robot's use of iconic gestures affected children's learning outcomes and levels of engagement.*

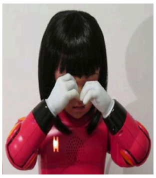
*Based on our survey of existing literature, we further presented ten outstanding questions in the research field, as well as four methodological suggestions. We were unable to address all of the outstanding questions in the present thesis, because several of them depend on future (technological) developments, because we only used one robot (NAO), and because we had to limit our scope (e.g., focusing on second language learning) for feasibility reasons. Our research (partly) addressed six of the outstanding questions (1, 2, 4, 5, 7, and 8). For example, the dataset of human-performed gestures that is presented in Chapter 5 can help improve the design of robot-performed gestures (Q1); in Chapter 6 we explore the effects of variation in the robot's gesturing behavior (Q4); in Chapters 3, 4, and 6 we investigate whether age influences the effects of robot-performed iconic gestures (Q7); and in Chapter 4 we look at spontaneous mirroring of the robot's gestures (Q8).*

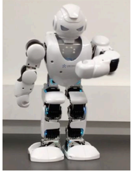
*Our methodological suggestions focused on replicability, external validity, the measurement instruments used, and connecting more deeply with related fields of research (e.g., human gesture studies). We follow these suggestions as well as possible in the upcoming chapters, which is why our studies take place in the field, the source code for the systems used in these studies is publicly available, and we provide elaborate documentation of the design and evaluation of the robot's gestures, as well as the measurement instruments used. This documentation also addresses outstanding questions 2 and 5, that relate to a need for more structure regarding the design and evaluation of robot-performed gestures.*

## 2.A Overview of robots used in research


(a) A100

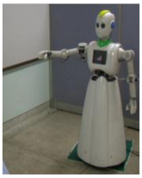
(b) Actroid-SIT

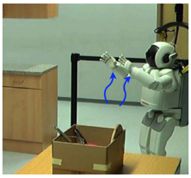

(c) ALICE

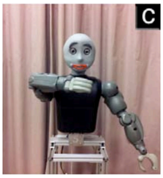

(d) Alpha (1)

(e) Alpha (2)

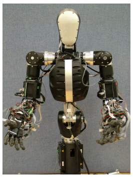
(f) AMI

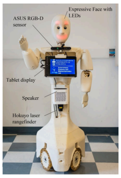

(g) AMIET


(h) ASIMO



(i) Bandit

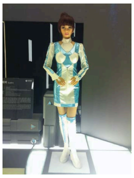

(j) BERTI



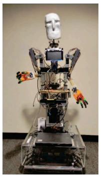
(k) Casper

(l) DARwIn-OP2

(m) Erica


(n) EveR-4E

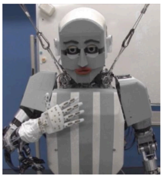
(o) iCub

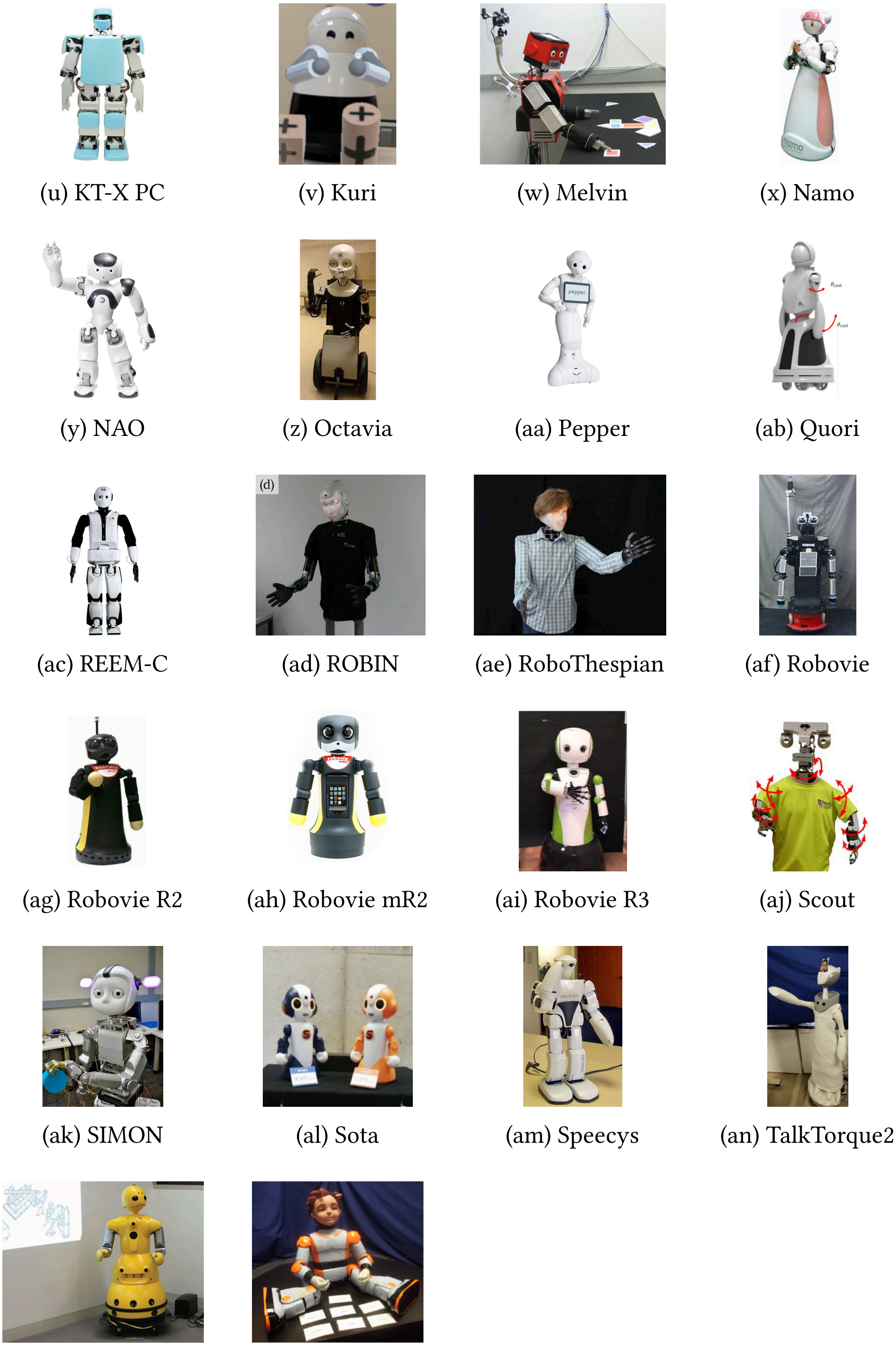

(p) Jeeves

(q) Justin

(r) KASPAR

(s) KHR2-HV


(t) KOBIAN



(u) KT-X PC (v) Kuri (w) Melvin (x) Namo

(y) NAO (z) Octavia (aa) Pepper (ab) Quori

(ac) REEM-C (ad) ROBIN (ae) RoboThespian (af) Robovie

(ag) Robovie R2 (ah) Robovie mR2 (ai) Robovie R3 (aj) Scout

(ak) SIMON (al) Sota (am) Speecys (an) TalkTorque2

(ao) Wakamaru (ap) Zeno R50

Figure 2.16: Robots used in the research covered in this literature review.

## 2.B Table of articles describing planning approaches

| Reference | Theme | Robot | Gesture type | Method | Validation |
|---|---|---|---|---|---|
| Admoni et al. (2016) | Selection | NAO | Deictic | Visual attention | U (N = 46) |
| Aly and Tapus (2012) | Synthesis | NAO | Emblematic | Data-driven | C |
| Aly and Tapus (2016) | Adjusting parameters | NAO | Iconic, metaphoric | Personality recognition | U (N = 21) |
| Aly and Tapus (2020) | Adjusting parameters | ALICE | Emblematic | Selection based on keywords | U (N = 60) |
| Augello and Pilato (2019) | Selection, timing | NAO | Iconic, metaphoric | Text analysis, WordNet | N |
| Augustine et al. (2020) | Selection, parameters | Erica | Emblematic | Selecting from a set | U (N = 34) |
| Bourguet et al. (2020a) | Selection | Pepper | Deictic, metaphoric | Real-time affect monitoring | U (N = 16) |
| Bremner et al. (2009) | Selection | BERTI | Beat | Rule-based | U (N = ?) |
| Claret et al. (2017) | Adjusting parameters | Pepper | Emblematic | Jacobian null space | U (N = 30) |
| Ghosh et al. (2019) | Selection | NAO | Emblematic | Random forest | U (N = 24) |
| Gielniak et al. (2011) | Adjusting parameters | SIMON | Emblematic | Task-aware variation | C |
| González et al. (2019) | Adjusting parameters | Pepper | Emblematic | Neural network | N |
| Gulzar and Kyrki (2015) | Adjusting parameters | NAO | Deictic | Probabilistic model | C |
| Hato et al. (2010) | Adjusting parameters | Robovie | Deictic | Region cognition model | U (N = 12) |
| Holroyd et al. (2011) | Adjusting parameters | Melvin | Deictic | Policies based on environment | U (N = 29) |
| Hua et al. (2019) | Synthesis | Pepper | Emblematic | Sequence to sequence | C |
| Huang and Mutlu (2014) | Selection | Wakamaru | Deictic, beat, iconic, metaphoric | Dynamic Bayesian Network | C, U (N = 29) |
| Hwang et al. (2020) | Selection | NAO | Deictic, metaphoric | Neural network | U (N = 20) |
| Ishi et al. (2018) | Selection | Erica | All | Probabilistic models | U (N = 20) |
| Ishi et al. (2019) | Timing | Erica | Deictic, beat, iconic, metaphoric | Overwriting/interpolating motions | U (N = 31) |
| Ishi et al. (2020) | Adjusting parameters | Erica | Deictic | Rule-based, dynamic speed and hold | U (N = 36) |
| Jung et al. (2004) | Adjusting parameters | AMI | Emblematic | Perception, motivation, memory | N |
| Jutharee and Maneewarn (2016) | Adjusting parameters | Namo | Emblematic | Genetic algorithm optimization | C |
| H. Kim et al. (2008) | Adjusting parameters | AMIET | Deictic, beat, emblematic | Modifying size, speed, frequency | U (N = 31) |
| Ko et al. (2020) | Synthesis | Pepper | Emblematic | Sequence to sequence | C |

| Reference | Theme | Robot | Gesture type | Method | Validation |
|---|---|---|---|---|---|
| Kondo et al. (2012) | Adjusting parameters | Actroid-SIT | Deictic, iconic, emblematic | Reconfigurable motion database | C, U (N = 42) |
| Kraemer et al. (2016) | Adjusting parameters | NAO | Emblematic | Audio analysis | C, U (N = 17) |
| Le et al. (2011) | Adjusting parameters | NAO | All but adaptor | Data-driven | N |
| Lemme et al. (2013) | Adjusting parameters | iCub | Deictic | Neural networks | C |
| S. Lim et al. (2009) | Selection | Speecys | Emblematic | Behavior network | U (N = 10) |
| A. Lim et al. (2011) | Adjusting parameters | NAO | Emblematic | Mapping voice elements to movement | U (N = 49) |
| Liu et al. (2017) | Adjusting parameters | Robovie R2 | Deictic | Computational model | U (N = 33) |
| Marmpena et al. (2019) | Synthesis, parameters | Pepper | Emblematic | Variational autoencoder | N |
| Marmpena et al. (2020) | Synthesis, parameters | Pepper | Emblematic | Conditional variational autoencoder | U (N = 20) |
| Mlakar et al. (2013) | Selection, timing | iCub | All but beat | TTS with prosodic/linguistic feats. | U (N = 30) |
| Ng-Thow-Hing et al. (2010) | Selection, timing, parameters | ASIMO | Deictic, beat, iconic, metaphoric | Probablistic model | U (N = 29) |
| Ondáš et al. (2017) | Selection | NAO | All | Rule-based | U (N = 30) |
| Ondras et al. (2020) | Synthesis | Pepper | Emblematic | Neural network | C, U (N = 63) |
| Paplu et al. (2020) | Selection, parameters | ROBIN | Emblematic | Monitor robot's internal state | U (N = 20) |
| Pérez-Mayos et al. (2020) | Selection | REEM-C | Beat, emblematic | Part of speech grammar & prosody | U (N = 50) |
| Peters et al. (2019) | Adjusting parameters | NAO | Unknown | Modifying head tilt, expansiveness | U (N = 31) |
| Prajod and Hindriks (2020) | Adjusting parameters | NAO | Deictic, metaphoric | Modifying amplitude, repetition, etc. | U (N = 396) |
| Ranatunga et al. (2015) | Adjusting parameters | Zeno | Emblematic | Dynamic movement primitives | N |
| Rodriguez et al. (2018) | Adjusting parameters | NAO | All | Sensor-based classification | N |
| Rodriguez et al. (2019) | Synthesis | NAO | Adaptor | Generative adversarial networks | C |
| Salem et al. (2013b) | Timing | ASIMO | Iconic, metaphoric | Phonological encoding & sensors | N |
| Shi et al. (2010) | Selection, timing, parameters | Multiple | Deictic, beat, iconic | Annotations in text | N |
| Shimazu et al. (2018) | Synthesis | Pepper | Iconic | Sequence to sequence | C |
| Stolzenwald and Bremner (2017) | Adjusting parameters | NAO | Emblematic | Modifying parameters | U (N = 99) |
| Sunardi and Perkowski (2020) | Selection, parameters | Jeeves | All | Probabilistic | N |
| Szafir and Mutlu (2012) | Adjusting parameters | Wakamaru | Deictic, beat, metaphoric | Brain-computer interface (EEG) | U (N = 30) |
| Tay and Veloso (2012) | Selection, parameters | NAO | Deictic, beat, iconic, metaphoric | Gesture primitives, rule-based | N |
| Tuyen et al. (2020a) | Synthesis | Pepper | All | Generative adversarial network | C |

| Reference | Theme | Robot | Gesture type | Method | Validation |
|---|---|---|---|---|---|
| Tuyen et al. (2020b) | Synthesis | Pepper | All | Generative adversarial network | C |
| Tuyen et al. (2021) | Adjusting parameters | Pepper | Emblematic | Interlocutor’s gesturing behavior | U (N = 109) |
| Valenti et al. (2020) | Adjusting parameters | NAO | Emblematic | Emotion recognition & mass spring | U (N = 25) |
| van de Perre et al. (2018) | Adjusting parameters | NAO, Justin | Deictic | ‘Blending’ emotional motion 1 limb | N |
| Viergutz et al. (2014) | Adjusting parameters | NAO | Emblematic | Selection & modifying parameters | N |
| J. Xu et al. (2013) | Adjusting parameters | NAO | Deictic, emblematic | Modifying parameters | U (N = 25) |
| Yoon et al. (2019) | Synthesis | NAO | Deictic, beat, iconic, metaphoric | End-to-end neural network | U (N = 46) |
| Yu and Tapus (2020) | Synthesis | Pepper | All | Generative adversarial networks | C |

Table 2.2: Information on articles included in the results describing planning approaches. Validation methods (of the planning approach): U = user study; C = computational; N = none. The full table is available in the supplementary materials: https://osf.io/uj9fq/.

## 2.C Table of articles describing experimental studies

| Reference | Theme | Robot | Gesture type | Location | Sample | Group | # Sess. | Results |
|---|---|---|---|---|---|---|---|---|
| Abdul Jalil et al. (2012) | C | NAO | Deictic, metaphoric, emblematic | Lab | Unknown | 1 | ? | None |
| Admoni et al. (2016) | P, T | NAO | Deictic | Lab | Adults (N = 46) | 1 | 1 | Pos. |
| Ahmad et al. (2016a) | T | NAO | Emblematic | Field | Children (10–12 y/o; N = 12) | 1 | 3 | Pos. |
| Ahmad et al. (2016b) | E, T | NAO | Emblematic | Field | Adults (N = 8) | 1 | 1 | Pos. |
| Ahn et al. (2013) | P | EveR-4E | Unknown | Field | Unknown | P | 1 | None |
| Aizawa and Umemuro (2021) | E | Sota | Emblematic | Field | Adults (N = 2162) | P | 1 | Neg. |
| Akalin et al. (2013) | S | NAO | Sign language | Lab | Adults (N = 14) | 1 | 1 | Mixed |
| Ali and Williams (2020) | T | NAO | Deictic | Lab | Adults (N = 8) | 1 | 1 | Pos. |
| Aly and Tapus (2015) | P | ALICE | Emblematic | Lab | Adults (N = 60) | 1 | 1 | Mixed |
| Aly and Tapus (2016) | P | NAO | Iconic, metaphoric | Lab | Adults (N = 21) | 1 | 1 | Pos. |
| Aly and Tapus (2020) | P | ALICE | Emblematic | Lab | Adults (N = 60) | 1 | 1 | Pos. |
| Asselborn et al. (2017) | P, E, T | NAO | Adaptor | Field | Children (5 y/o; N = 20) | 1 | 1 | Pos. |
| Augustine et al. (2020) | P | Erica | Emblematic | Unknown | Adults (N = 34) | 1 | 1 | Pos. |
| Bao and Cuijpers (2017) | C, P | NAO | Deictic | Lab | Adults (N = 70) | 1 | 1 | None |
| Bennewitz et al. (2005) | C, P, T | Alpha (2) | Deictic | Lab | Unknown | ? | 1 | Pos. |
| Bourguet et al. (2020a) | E | Pepper | Deictic, metaphoric | Lab | Unknown (N = 16) | ? | 1 | Pos. |
| Bourguet et al. (2020b) | E | Pepper | Deictic, beat, iconic, metaphoric | Online | Adults (N = 56) | 1 | 1 | Neg. |
| Bremner et al. (2011) | E, T | BERTI | Deictic, beat, iconic | Lab, Field | Adults (N = 130) | >= 3 | 1 | Pos. |
| Bremner and Leonards (2015a) | C | NAO | Iconic | Lab | Adults (N = 22) | 1 | 1 | Pos. |
| Bremner and Leonards (2015b) | C | NAO | Beat | Lab | Adults (N = 22) | 1 | 1 | Neg. |
| Bremner et al. (2016) | P | NAO | Unknown | Online | Adults (N = 143) | 1 | 1 | Pos. |
| Burns et al. (2018) | P | ROBOTIS-OP2 | Emblematic | Lab | Adults (N = 12) | 1 | 1 | Pos. |
| Cabibihan et al. (2012) | C | Scout | Iconic, emblematic | Online | Adults (N = 122) | 1 | 1 | Neg. |
| Carter et al. (2014) | P, E | A100 | Emblematic | Lab | Adults (N = 30) | 1 | 1 | Pos. |
| Chidambaram et al. (2012) | T | Wakamaru | Deictic, beat, iconic, metaphoric | Lab | Adults (N = 32) | 1 | 1 | Pos. |

| Reference | Theme | Robot | Gesture type | Location | Sample | Group | # Sess. | Results |
|---|---|---|---|---|---|---|---|---|
| Claret et al. (2017) | P | Pepper | Emblematic | Lab | Adults (N = 30) | 1 | 1 | Pos. |
| De Carolis et al. (2019) | E, T | NAO | Emblematic | Lab | Children & adults (N = 10) | 1 | 1 | Pos. |
| Chapter 3 (de Wit et al., 2018) | E, T | NAO | Iconic | Field | Children (4–6 y/o, N = 61) | 1 | 1 | Pos. |
| Chapter 6 (de Wit et al., 2020) | E, T | NAO | Iconic | Field | Children (4–6 y/o, N = 94) | 1 | 1 | Pos. |
| DePalma et al. (2021) | T | NAO | Deictic | Online | Adults (N = 100) | 1 | 1 | Pos. |
| Deshmukh et al. (2018) | P | Pepper | Deictic, emblematic | Lab | Adults (N = 30) | >= 3 | 3 | None |
| Émond et al. (2020) | P | NAO | Emblematic | Field | Adults & children (N = 121) | 2 | 1 | Pos. |
| Ghosh et al. (2019) | P | NAO | Emblematic | Lab | Adults (N = 24) | >= 3 | 1 | None |
| Gielniak and Thomaz (2012) | P, T | SIMON | Iconic | Lab | Adults (N = 122) | 1 | 1 | Pos. |
| Goto et al. (2020) | P | Sota | Deictic, emblematic | Field | Adults (N = 78) | 1 | 1 | None |
| Groechel et al. (2019) | P, T | Kuri | Deictic, metaphoric, emblematic | Lab | Adults (N = 34) | 1 | 1 | Pos. |
| Ham et al. (2015) | P, T | NAO | Iconic, metaphoric | Lab | Adults (N = 64) | 2 | 1 | Pos. |
| Hasegawa et al. (2010) | P, T | KHR2-HV | Deictic | Lab | Adults (N = 74) | 1 | 1 | Pos. |
| Hayes et al. (2013) | C | NAO | Iconic | Lab | Adults (N = 54) | 1 | 1 | Neg. |
| Hsieh et al. (2020) | P | Pepper | Emblematic | Field | Adults (N = 9) | 1 | 2 | None |
| Huang and Mutlu (2013) | T | Wakamaru | Deictic, beat, iconic, metaphoric | Lab | Adults (N = 32) | 1 | 1 | Pos. |
| Huang and Mutlu (2014) | P, T | Wakamaru | Deictic, beat, iconic, metaphoric | Lab | Adults (N = 29) | 1 | 1 | Mixed |
| Iio et al. (2011) | C, P | Robovie R2 | Deictic | Lab | Adults (N = 18) | 1 | 1 | Mixed |
| Isaka et al. (2018) | C | Sota | Adaptor | Lab | Unknown (N = 12) | ? | 1 | Pos. |
| Jouaiti and Henaff (2019) | E, T | Pepper | Emblematic | Lab | Adults (N = 15) | 1 | 1 | Mixed |
| Kanda et al. (2002) | C | Robovie | Deictic, emblematic | Lab | Adults (N = 36) | 1 | 1 | Pos. |
| Kashii et al. (2016) | C | NAO | Emblematic | Lab | Adults (N = 7) | 1 | 1 | Pos. |
| Kaushik and Simmons (2021) | P | Quori | Emblematic | Online | Adults (N = 145) | 1 | 1 | None |
| Kawaguchi et al. (2016) | P | TalkTorque 2 | Deictic | Lab | Adults (N = 12) | 1 | 1 | Pos. |
| H. Kim et al. (2008) | P | AMIET | Deictic, beat, emblematic | Lab | Adults (N = 31) | 1 | 1 | None |
| A. Kim et al. (2012) | P, T | NAO | Unknown | Lab | Adults (N = 63) | 1 | 1 | Mixed |
| A. Kim et al. (2013) | P, E | NAO | Unknown | Lab | Adults (N = 27) | 1 | 3 | Pos. |
| K. Kim et al. (2017) | E | RoboThespian | Unknown | Field | All (N = 7685) | P | >= 1 | None |

| Reference | Theme | Robot | Gesture type | Location | Sample | Group | # Sess. | Results |
|---|---|---|---|---|---|---|---|---|
| Kose et al. (2012) | S | NAO | Sign language | Online | Adults & Children (N = 113) | 1 | 1 | None |
| Kose et al. (2015) | S | Multiple | Sign language | Lab | Adults & Children (N = 39) | 1 | 1 | None |
| N. Lee et al. (2017) | T | NAO | Deictic, emblematic | Field | Adults (N = 118) | 1 | 1 | Pos. |
| Li et al. (2015) | P | NAO | Beat | Online | Adults (N = 60) | 1 | 1 | Mixed |
| Ligthart et al. (2019) | P | NAO | Emblematic | Field | Children (8–11 y/o, N = 75) | 1 | 1 | None |
| Ligthart et al. (2020) | P | NAO | Iconic | Field | Children (8–10 y/o, N = 27) | 1 | 1 | Pos. |
| Liles et al. (2017) | P, T | NAO | Deictic, metaphoric | Lab | Adults (N = 30) | 1 | 1 | Pos. |
| S. Lim et al. (2009) | T | Speecys | Emblematic | Lab | Adults (N = 10) | 1 | 1 | Pos. |
| Lohse et al. (2014) | P, T | NAO | Deictic | Lab | Adults (N = 32) | 1 | 1 | Mixed |
| Marmpena et al. (2020) | P | Pepper | Unknown | Lab | Adults (N = 20) | 1 | 1 | Pos. |
| Meena et al. (2012) | C, P, E | NAO | Beat, emblematic | Field | Adults (N = 12) | 1 | 1 | Pos. |
| Mikata et al. (2019) | P | Multiple | Deictic, iconic, metaphoric | Lab | Adults (N = 17) | 1 | 1 | Pos. |
| Moro et al. (2019) | P, E, T | Casper | Deictic, beat | Field | Elderly (N = 6) | 1 | 6 | Pos. |
| Moshkina et al. (2014) | E | Octavia | All but adaptor/sign | Field | Adults & children (N = 4222) | P | 1 | Pos. |
| Muto et al. (2009) | P | Wakamaru | Deictic | Lab | Adults & elderly (N = 36) | 1 | 1 | None |
| Nalin et al. (2012) | C | NAO | Emblematic | Field | Children (5–12 y/o, N = 13) | 1 | 3 | None |
| Okuno et al. (2008) | P, T | Robovie | Deictic | Lab | Adults (N = 21) | 1 | 1 | Pos. |
| Ondáš et al. (2017) | P | NAO | All but sign | Lab | Adults (N = 30) | 1 | 1 | Pos. |
| Ono et al. (2001) | P, T | Robovie | Deictic | Lab | Adults (N = 30) | 1 | 1 | None |
| Park et al. (2011) | P | Multiple | Unknown | Lab | Adults (N = 40) | 1 | 1 | Pos. |
| Peters et al. (2019) | P | NAO | Unknown | Lab | Adults (N = 31) | 1 | 1 | None |
| Pollmann et al. (2020) | P | Pepper | Unknown | Lab | Adults (N = 12) | 1 | 1 | Pos. |
| Prajod and Hindriks (2020) | P | NAO | Deictic, emblematic | Online | Adults (N = 396) | 1 | 1 | Pos. |
| Rehm et al. (2016) | P | NAO | Emblematic | Lab | Adults & elderly (N = 27) | 1 | 1 | None |
| Riek et al. (2010) | C | BERTI | Emblematic | Lab | Adults (N = 16) | 1 | 1 | None |
| Robins et al. (2008) | C | KASPAR | Emblematic | Field | Children (N = 18) | 1 | 1 | None |
| Sakamoto et al. (2005) | P | Robovie | Deictic | Lab | Adults (N = 50) | 1 | 1 | Pos. |
| Salem et al. (2011) | P | ASIMO | Deictic, iconic | Lab | Adults (N = 81) | 1 | 1 | Pos. |

| Reference | Theme | Robot | Gesture type | Location | Sample | Group | # Sess. | Results |
|---|---|---|---|---|---|---|---|---|
| Salem et al. (2012) | P | ASIMO | Deictic, iconic | Lab | Adults (N = 60) | 1 | 1 | Mixed |
| Salem et al. (2013a) | P, T | ASIMO | Deictic, iconic | Lab | Adults (N = 62) | 1 | 1 | Pos. |
| Salvador et al. (2015) | S | Zeno R-50 | Emblematic | Lab | Children, NT & ASD (N = 22) | 1 | 1 | Mixed |
| Sauppé and Mutlu (2014) | C | NAO | Deictic | Lab | Adults (N = 24) | 1 | 1 | None |
| Seo et al. (2014) | P | DARwIn-OP | Emblematic | Unknown | Unknown (N = 11) | 1 | 1 | Mixed |
| Seo et al. (2015) | P | DARwIn-OP | Emblematic | Unknown | Unknown (N = 21) | 1 | 1 | None |
| Shimazu et al. (2018) | P | Pepper | Iconic | Lab | Adults (N = 13) | 1 | 1 | Pos. |
| So et al. (2016) | S | NAO | Iconic, emblematic | Field | Children (N = 20) | 1 | 12 | Pos. |
| So et al. (2018a) | S | NAO | Iconic, emblematic | Field | Children (N = 13) | 1 | 8 | Pos. |
| So et al. (2018b) | S | NAO | Iconic, emblematic | Field | Children (N = 45) | 1 | 4 | Pos. |
| So et al. (2019a) | S | NAO | Iconic, emblematic | Field | Children (N = 24) | 1 | 9 | Pos. |
| So et al. (2019b) | S | NAO | Iconic, emblematic | Field | Children (N = 23) | 1 | 4 | Pos. |
| St. Clair et al. (2011) | C | Bandit | Deictic | Lab | Adults (N = 40) | 1 | 1 | Mixed |
| Stolzenwald and Bremner (2017) | C | NAO | Unknown | Online | Adults (N = 61) | 1 | 1 | Pos. |
| Sugiyama et al. (2007) | P | Robovie | Deictic | Lab | Adults (N = 30) | 1 | 1 | Pos. |
| Szafir and Mutlu (2012) | E, T | Wakamaru | Deictic, beat, metaphoric | Lab | Adults (N = 30) | 1 | 1 | Pos. |
| Taheri et al. (2020) | S | NAO | Emblematic | Lab | Children (N = 40) | 1 | 1 | Neg. |
| Tahir et al. (2020) | C, P | NAO | Emblematic | Lab | Adults (N = 20) | 1 | 1 | Pos. |
| Thepsoonthorn et al. (2021) | P | NAO | Deictic, emblematic | Lab | Adults (N = 20) | 1 | 1 | Pos. |
| Tielman et al. (2014) | P | NAO | Emblematic | Field | Children (8–10 y/o, N = 18) | 1 | 1 | Pos. |
| Trovato et al. (2013) | P | KOBIAN | Emblematic | Lab | Adults (N = 36) | 1 | 1 | None |
| Tsiourti et al. (2017) | P | Multiple | Emblematic | Online | Adults (N = 170) | 1 | 1 | Pos. |
| Tuyen et al. (2021) | P | Pepper | Emblematic | Lab | Adults (N = 286) | 1 | 1 | Pos. |
| Valenti et al. (2020) | C, P | NAO | Emblematic | Online | Adults (N = 161) | 1 | 1 | Pos. |
| van den Berghe et al. (2021a) | P | NAO | Deictic, iconic | Field | Children (5 y/o, N = 104) | 1 | 7 | None |
| van Dijk et al. (2013) | T | NAO | Iconic | Unknown | Elderly (N = 19) | 1 | 1 | Pos. |
| Vogt et al. (2017b) | P, T | NAO | Deictic, iconic, emblematic | Unknown | Adults, learning imp. (N = 33) | 1 | 1 | Pos. |
| Vogt et al. (2019) | T | NAO | Deictic, iconic | Field | Children (5 y/o, N = 194) | 1 | 7 | None |

| Reference | Theme | Robot | Gesture type | Location | Sample | Group | # Sess. | Results |
|---|---|---|---|---|---|---|---|---|
| X. Wang et al. (2014) | C | NAO | Deictic | Lab | Adults & children (N = 36) | 1 | 1 | None |
| Wicke and Veale (2020) | P | NAO | Deictic, emblematic | Online | Adults (N = 121) | 1 | 1 | None |
| Wu et al. (2017) | E | NAO | Deictic, iconic, metaph., emblem. | Field | Children (14–15 y/o, N = 60) | 1 | 1 | None |
| Wykowska et al. (2015) | P | NAO | Deictic | Lab | Adults (N = 18) | 1 | 1 | None |
| J. Xu et al. (2014) | P, T | NAO | Unknown | Field | Adults (N = 34) | >= 3 | 1 | Pos. |
| J. Xu et al. (2015a) | P | NAO | Unknown | Lab | Adults (N = 66) | 1 | 1 | Pos. |
| J. Xu et al. (2015b) | P, T | NAO | Emblematic | Lab | Adults (N = 36) | 1 | 1 | Pos. |
| K. Xu (2019) | P | Alpha (1) | Emblematic | Lab | Adults (N = 110) | 1 | 1 | Mixed |
| Yadollahi et al. (2018) | T | NAO | Deictic | Field | Children (6–7 y/o, N = 16) | 1 | 2 | Mixed |
| Yamazaki et al. (2016) | C | Custom | Emblematic | Lab | Adults (N = 118) | 1 | 1 | Neg. |
| Zhang and de Haas (2020) | T | NAO | Metaphoric | Lab | Adults (N = 21) | 1 | 1 | Mixed |
| Zhao et al. (2016) | C | Multiple | Deictic | Online | Adults (N = 1648) | 1 | 1 | None |
| Zhao and Malle (2019) | C | Multiple | Deictic | Online | Adults (N = 1690) | 1 | 1 | Pos. |
| M. Zheng and Meng (2012) | C | NAO | Emblematic | Lab | Adults (N = 16) | 1 | 1 | Neg. |
| Z. Zheng et al. (2016) | S | NAO | Emblematic | Lab | Children (3–4 y/o, N = 16) | 1 | 4 | Pos. |
| M. Zheng et al. (2019) | C | NAO | Emblematic | Lab | Adults (N = 64) | 1 | 1 | Neg. |

Table 2.3: Information on articles included in the results describing experimental studies. Themes: C = communicative purposes; P = perception; E = engagement; T = task performance; S = special needs. Group size P = passersby. *Group* indicates the number of people the robot interacted with (1, 2, >= 3, or any number of passersby). *Sessions* describes the number of sessions that one participant had with the robot. *Results* indicates whether the study found a positive, negative, or no effect of the robot's use of gestures on the included outcome measures. No effect could also indicate a neutral effect, where the gestures changed an outcome measure, such as the perception of the robot (but not for better or for worse). The full table is available in the supplementary materials: https://osf.io/uj9fq/.

3

## Chapter 3

# A First Study into Robot-Performed Iconic Gestures to Support Second Language Learning

*This chapter is based on:* de Wit, J., Schodde, T., Willemsen, B., Bergmann, K., de Haas, M., Kopp, S., Krahmer, E. & Vogt, P. (2018, February). 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).[2] *Open practices:* The materials are available at https://github.com/l2tor/animalexperiment. The experiment was not preregistered.

## Abstract

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

## 3.1 Introduction

Robots show great potential in the field of education (Mubin et al., 2013). Embodied agents in the form of humanoid robots, in particular, may deliver educational content for various subjects in ways similar to human tutors. The main advantage of using such a robot compared to traditional learning tools is its physical presence in the referential world of the learner (Leyzberg et al., 2012). The human-like appearance and presence in the physical environment may facilitate interactions that are, to some extent, similar to the ways in which human teachers would communicate with their students. Care should be taken, however, to design for the correct amount of social behavior, so as to avoid distracting students from the task at hand (Kennedy et al., 2015).

When designing such interactions, we can draw upon ways in which human teachers give contingent support to students in their learning activities. For instance, particularly in one-on-one tutoring situations, teachers tend to adjust the pace and difficulty of learning tasks based on the past development and current skill set of the student (van de Pol et al., 2010). For example, teachers may help by scaffolding, taking the initial knowledge base as a starting point and trying to optimize the learning gain by choosing the hardest task to perform that still lies within the zone of proximal development (Vygotsky, 1978) of the student.

The use of gestures that coincide with speech is another way for teachers to provide scaffolding, particularly when the concepts which the gestures refer to are not yet mastered by the student (Alibali & Nathan, 2007). For instance, when teaching a second language (L2), gestures can help to ground an unknown word in the target language by linking it iconically or indexically to a real world concept. Such a facilitating effect on word learning has been found for imitating gestures of a virtual avatar (Bergmann & Macedonia, 2013). However, it is an open question if the embodied presence of a robot can be exploited to support language learning through a robot's gesturing, and if so, what kind of gestures would have a positive impact.

In this chapter, we present the results of an experiment conducted to explore how these two tools for scaffolding the learning of language — choosing the task that yields the greatest potential learning gain for a particular student and the use of appropriate co-speech gestures — carry over to a humanoid robot. Both were combined in one study to better estimate what the relative importance of the respective techniques is, while keeping all other factors constant, and to find out whether the benefits of the two strategies can potentially reinforce or impede each other. The techniques were implemented and tested in a one-on-one tutoring system where children, four to six years old, play a game with a robot to learn an L2. In the next section, we briefly present the approaches taken to realize the adaptive tutoring along with co-speech gesturing of the robot. We then describe the experimental methodology, before reporting and discussing the results obtained.

## 3.2 Background

### 3.2.1 Adaptive Bayesian Knowledge Tracing

A robot tutor that personalizes the learning experience for individual students has been shown to have a positive effect on performance (Leyzberg et al., 2014). This robot is also perceived as smarter or more intelligent and less distracting or annoying. In order to simulate the way human tutors tailor learning activities and difficulty levels to a particular student, an adaptive tutoring system would have to measure and track the knowledge level of the student. Often the knowledge is traced skill-wise, where in the case of language learning, the mastery of particular words or phrases in the target language is represented probabilistically (e.g., Gordon & Breazeal, 2015). This approach yields promising results, but it lacks flexibility because of the need to define domain-specific distance metrics to choose the next skill. Others have used Dynamic Bayesian Networks to represent the learner's knowledge about a skill, conditioned on the past interaction and taking into account skill interdependencies (Käser et al., 2014). This approach requires detailed knowledge about the learning domain to model those interdependencies and their parameters. Recently, Spaulding et al. (2016) used a simpler approach based on Bayesian Knowledge Tracing (BKT) (Corbett & Anderson, 1994). The general BKT model consists of latent variables $S^t$ representing the extent to which the system believes a particular skill to be mastered by the student. The belief state of the system is updated based on observed variables $O^t$, which correspond to the result of a learning action (e.g., correctly or incorrectly answering a question), while accounting for possible cases of guessing *p(guess)* and slipping *p(slip)* during the answer process. It was shown that this model outperforms traditional approaches for tracing the knowledge state in learning interactions, and that it can be easily extended to, for example, incorporate the emotional state of a child. In previous work (Schodde et al., 2017), we have extended the basic BKT with action nodes to also model the tutor's decision-making based on current beliefs about

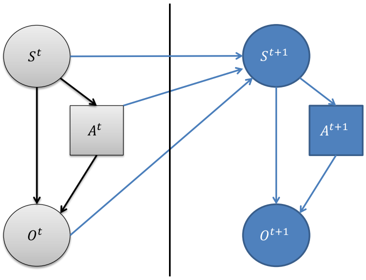


Figure 3.1: Dynamic Bayesian Network for BKT taken from Schodde et al., 2017, with permission: with the current skill-belief the robot chooses the next skill $S^t$ and action $A^t$ for time step $t$ and observes $O^t$ as response from the user.

the student's knowledge state (see Figure 3.1). Additionally, we employed a latent variable $S$ that can attain discrete values for each skill, corresponding to six bins for the belief state (0%, 20%, 40%, 60%, 80%, 100%). This allows for quantifying the robot's uncertainty about a learner's skills as well as the impact of tutoring actions on future observations and skills.

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

The goal of this strategy is to create a feeling of flow which can lead to better learning results (Craig et al., 2004). It strives not to overburden the learner with tasks that would be too difficult nor to bore them with tasks that would be too easy, both of which may lead to disengagement and thus hamper the learning. Note that this approach is comparable to the vocabulary learning technique of *spaced repetition* as implemented, for instance, in the Leitner system (Leitner, 1972). The implementation of A-BKT used in the current study is identical to the one used previously in Schodde et al. (2017). However, it has not yet been evaluated with children nor in conjunction with other techniques that might affect action difficulty (such as gestures). Furthermore, its impact on student engagement has not been explored previously.

### 3.2.2 Gestures

Iconic gestures elicit a mental image that corresponds directly, either in form or execution, to the concept or action that is being described verbally at the same time (McNeill, 1985). For example, a flying bird could be depicted by stretching both arms sideways and moving them up and down. Studies have shown that iconic gestures, when performed by a human teacher, may aid the acquisition of L2 vocabularies (de Nooijer et al., 2013; Kelly et al., 2009; Macedonia et al., 2011; Tellier, 2008). Hald et al. (2016) provide an overview of how gestures can contribute to learning an L2. They propose that gestures might have a 'grounding' effect by linking existing perceptual and motor experiences to a new word. This is expected to result in a richer mental representation. Research by Rowe et al. (2013) shows that gender, language background, and level of experience in the native language (L1) influence the extent to which gestures can contribute to L2 learning. The positive effects of gestures hold true for younger students as well; in fact, gestures are suggested to be a crucial part of communication with children (Hostetter, 2011). It has also been shown that gestures help not only to acquire knowledge, but also to retain it over time (Cook et al., 2008).

Previous research has explored the use of gestures by virtual agents (e.g., Bergmann & Macedonia, 2013) and robots (e.g., van Dijk et al., 2013), finding similar, positive effects on memory performance when gestures are produced by an artificial embodied agent compared to a human tutor. While humans tend to spontaneously perform and time their gestures, they will often need to be manually designed and coordinated with speech for the robot. Due to its limited degrees of freedom, however, the robot is unable to perform motions with the same level of detail, finesse, and accuracy as a human. This may lead to a loss in meaning when human gestures are being translated directly to the robot, indicating a need for alternative gestures. As a concrete example, the SoftBank Robotics NAO robot that was used in this case is unable to move its three fingers individually, preventing it from performing pointing gestures or finger-counting. However, research suggests that iconic gestures are almost as comprehensible when performed by a robot, compared to a human (Bremner & Leonards, 2016).

## 3.3 Methodology

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

H1: There is a greater learning gain when target words are accompanied by iconic gestures during training, than in the case of not using gestures.

H2: There is a reduced knowledge decay when target words are accompanied by iconic gestures during training, than in the case of not using gestures.

H3: There is a greater learning gain when target words are presented in an adaptive order during training, based on the knowledge state of the child, than when target words are randomly introduced.

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

The experiment had a 2 (adaptive versus non-adaptive) x 2 (gestures versus no gestures) between-subjects design. In the two conditions with the adaptive tutoring strategy, the A-BKT system described in Section 3.2.1 was used to select the target word for each round, based on the believed knowledge state of the child. In practice, this meant that children would be presented with a particular target word more frequently if they had answered it incorrectly in the past, thereby changing the number of times each target word occurred during training, although each target word was guaranteed to occur at least once. Other conditions had a random selection, where each of the six target words would always be presented five times, in a randomized order, for a total of thirty rounds. In the gesture conditions, whenever a target word was introduced in the L2 it was accompanied by an iconic gesture (as shown in Figure 3.3). All conditions had the robot standing up and in "breathing" mode, which meant that it slowly shifted its weight from one leg to the other and had a slight movement in its arms to simulate breathing.

### 3.3.1 Participants

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

### 3.3.2 Materials

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

The SoftBank Robotics NAO robot was used, which was standing in front and slightly to the right of the child. After an experimenter had filled in the name of the child and pressed the start button, the experiment ran fully autonomously. Two experimenters were always present, where one would take care of getting the child from the classroom and explaining the procedure of the experiment, while the other would set up the system. To avoid having the child seek them out for feedback, the experimenters would announce that they would be occupied. The child was asked to sit on pillows, close to the tablet which was raised on a box and slightly tilted. Two cameras were used to record the interaction, one facing the front of the child and one at an angle from the side. The basic setup is shown in Figure 3.2, although it differed slightly between locations due to the layout of the rooms. In the condition with gestures every occurrence of the target word in L2, except when giving feedback, was accompanied by the matching iconic gesture (see Figure 3.3). The gesture was timed in such a way that the pronunciation of the target word would coincide with the stroke of the gesture, i.e., the accented phase that is most related to the meaning.

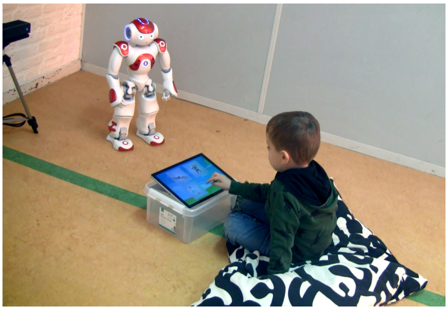


Figure 3.2: The setup for the experiments.

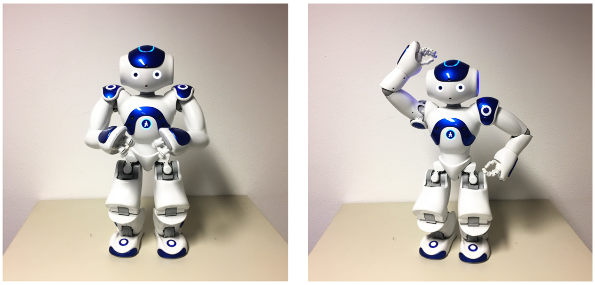


Figure 3.3: Examples of the stroke of two iconic gestures performed by the robot (taken from de Wit et al., 2017, with permission). Left: imitating a *chicken* by simulating the flapping of its wings; right: imitating a *monkey* by scratching head and armpit.

A perception study was conducted to evaluate the quality of the gestures (de Wit et al., 2017), where 14 participants were shown video recordings of all six gestures performed by the robot and then asked to indicate which out of the six target words corresponds to each particular recording. Based on the results of this study, each gesture was deemed to be sufficiently unique to distinguish between the six target words.

The adaptive tutoring system starts with medium (0.5) confidence for all target words, a value associated with two distractors during training. Each distractor is a false answer to a task, an image belonging to one of the five other target words. In the random conditions, since there is no knowledge tracing the difficulty was always set to medium (two distractors). The tablet was used to get input from the child, because speech recognition does not work reliably with children (Kennedy et al., 2017). This is also why only comprehension and not production of the target words is evaluated. An example of what the tablet screen would look like is shown in Figure 3.4. The images used during training belong to a different set of images than the ones used for the pre-test and post-tests. The set of images used during training matches the gesture that the robot performs related to the animals, for example the image of the horse for the training stage (shown in Figure 3.4) also includes a rider because the robot shows the act of riding a horse as a gesture. The image that was used during the tests did not include a rider and the horse is standing still, facing the opposite direction (shown in Figure 3.5). In addition to changing the pose or context of the animals, colors also varied. Together with having a recorded voice in the tests instead of the robot's synthesized speech, this aims to verify whether children learn how the English words map to the concepts of the animals and their matching Dutch

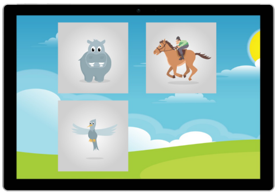


Figure 3.4: The tablet during training, showing images corresponding to the target word and two distractors.

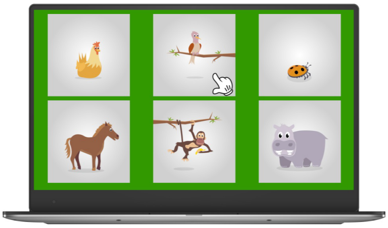


Figure 3.5: The pre-test and post-tests on a laptop, using a recorded voice and a different set of images from those on the tablet.

words, rather than to one specific image.

### 3.3.3 Procedure

Prior to partaking in the experiment, participants were introduced to the robot during a group introduction. This approach is inspired by the work of Vogt et al. (2017a) with the intention of lowering the anxiety of children in subsequent one-on-one interactions with the robot. The introduction consisted of a description of what the robot is like, including a background story and how it is similar to humans in some respects, and different in others. Together with the children (and sometimes teachers and experimenters) the robot performed dances, after which all children were presented with the opportunity to shake the robot's hand before putting it to bed. Introductory sessions were scheduled several days before the first participant was to take part in the experiment, allowing time for the children to process these new impressions.

Before starting the tutoring interaction, a pre-test was administered to gauge the level of prior knowledge with respect to the animal names in the L1 (Dutch) and L2 (English). This test was administered on a laptop, where images of all six animals were randomly positioned on the screen. A recording of a (bilingual) native speaker pronouncing one of the six animal names was played, after which the child was asked to click the corresponding image on the screen (Figure 3.5). This was done for all six target words, first in Dutch and then in English.

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

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

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

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

After the robot had "spied" an animal, a corresponding image was shown on the tablet along with a number of distractor images (Figure 3.4). The child was then asked to pick the image that matched the animal name that the robot had spied. The number of distractors was determined by the difficulty level of the round, which in the case of the adaptive conditions depended on the confidence that the system had in that the child knew this particular target word. A low confidence resulted in only one distractor, while a high confidence had three distractors.

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

### 3.3.4 Analysis

Immediate learning gain was measured as the difference between the number of correct answers on the post-test, administered directly after the training stage, and the number of correct answers on the pre-test, taken prior to the tutoring interaction. Test scores were always between 0 and 6 because each target word was asked once in the L2. The post-test was administered once more, (at least) one week after the experiment. We then looked at the difference between this delayed test and the pre-test for long-term learning gain. Finally, we took the difference between the delayed test and the immediate post-test as a measure of knowledge decay. The design of these tests is described in more detail in Section 3.3.2.

Children’s tasks during training were of varying task difficulty in the adaptive tutoring condition, with one to three distractor images. To account for these differences, as well as to allow a comparison with the post-test results (five distractor images), we mapped binary task success (1: correct response; 0: incorrect response) onto the span between 0.0 and 1.0 by subtracting a value of 0.2 for each of the potential five distractor images that was not provided, which would, for example, result in a score of 0.6 for a correct response in a task with three distractors. The total score during training was then divided by the number of rounds (30), resulting in a training performance value between 0.0 and 1.0 (Figure 3.6).

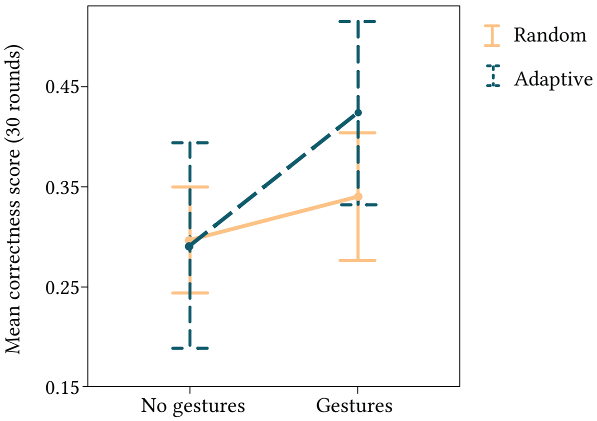


Figure 3.6: Interaction effects of gesture use and training strategy. The error bars are +/- 1 SD.

## 3.4 Results

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

To investigate the effects of the different conditions on training performance, a two-way ANOVA was carried out with tutoring strategy (adaptive versus non-adaptive) and the use of gestures (gestures versus no gestures) as independent variables and performance during training as the dependent variable (Figure 3.6). As described in Section 3.3.4, these scores are weighted by the number of distractors present and divided by 30 rounds, resulting in a value between 0.0 and 1.0. For the 30 rounds of training there was a main effect of gesture use, $F(1, 57) = 18.23, p < .001$, $\eta_p^2 = .24$, such that training with gestures led to higher score ($M = .38, SD = .09$) than learning without gestures ($M = .29, SD = .08$). Children in the adaptive condition achieved a higher score ($M = .36, SD = .12$) than children in the non-adaptive condition ($M = .32, SD = .06$), but the effect of tutoring strategy was not significant, $F(1, 57) = 3.62, p = .06$, $\eta_p^2 = .06$. There was a significant interaction effect between use of gestures and tutoring strategy, $F(1, 57) = 4.72, p = .03$, $\eta_p^2 = .08$. Without gesture use, there was no significant difference between tutoring strategies. When gestures were present, however, children in the adaptive condition turned out to perform better than those in the non-adaptive condition. Hence, children's learning outcome was best when gesture use and adaptive training were combined.

Another two-way ANOVA was carried out to measure learning gain, with the difference score between the post-test results and the pre-test results as the dependent variable (Figure 3.7). There was no significant effect of tutoring strategy, $F(1, 57) < .01$[2], $p = .95$, $\eta_p^2 < .001$, or use of gestures, $F(1, 57) = 1.53, p = .22$, $\eta_p^2 = .03$. These results do not support H1 and H3 (greater learning gains when gestures and adaptive tutoring are used). The same two-way ANOVA with the difference score between results of the delayed post-test and the pre-test also did not give a significant effect

[2] The original article reports this value as $< .001$, but $F$ was in fact .004.

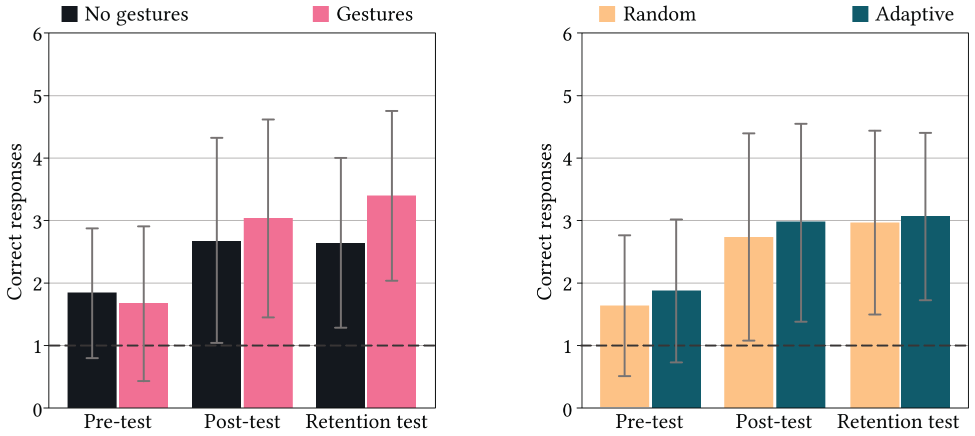


Figure 3.7: Test scores for the gesture vs no gesture conditions (left) and the adaptive vs random conditions (right). The error bars are +/- 1 SD.

of tutoring strategy, $F(1, 57) = .36, p = .55, \eta_p^2 = .006$, but there was a significant effect for use of gestures, $F(1, 57) = 6.11, p = .02, \eta_p^2 = .097$, indicating that the learning gain between pre-test and delayed post-test was greater when gestures were used during training ($M = 1.70, SD = 1.56$) than when no gestures were used ($M = 0.81, SD = 1.25$). Although this does not fully support H1 or H2, it does show a long-term learning gain when gestures are used during learning. No interaction effect was found, $F(1, 57) = .04, p = .84, \eta_p^2 \leqslant .001$.

### 3.4.1 Evaluation of engagement

The engagement of the children during the training stage with the robot was examined to find out whether children became more disengaged with the tutoring tasks toward the end of the thirty rounds, and whether the application of an adaptive tutoring strategy and gestures would influence the change in engagement levels. This was done by asking 18 adult participants, without specific training in working with children, to rate video clips (without audio) of the children interacting with the robot. The choice for conducting a perception study with adults using video recordings of the experiment was made for two reasons: so that the training would not have to be interrupted for questions regarding the experience, thereby potentially influencing the engagement, and because it is difficult for children of a young age to reflect upon their experiences and verbalize these thoughts (Markopoulos et al., 2008). For each child, one clip was taken from the fifth round of training and one clip from the twenty-fifth round, to get observations that are close to the beginning and end of the training, but far enough from these actual moments to avoid short bursts of engagement when children realize the experiment is starting or finishing. The clips start right after the robot finishes introducing the task, i.e., the point at which the turn switches to the child to provide an answer. All clips then run for five seconds. One child that was excluded from the previous analysis because delayed post-test results were missing, was included for this part of the evaluation. However, data from one other child was missing, making the number of stimuli 122 (61 children, two clips each), with 14 to 16 children in each condition. Participants in the evaluation were asked to rate all 122 clips, randomly presented to them, on a scale from 1 (completely disengaged) to 7 (completely engaged). As a practice round, two clips of a child that was not included in the main experiment were presented, where one example was clearly engaged and the other was clearly not engaged. After this practice round, participants were told which features from the examples showed engagement (i.e., rapid response to the question, upright body posture, displaying joy after answering the question) and disengagement (i.e., slower response to the question, supporting the head by leaning on the arms, showing less interest in the task).

For each participant, the ratings were averaged over all children belonging to the same experimental condition, resulting in a total of eight average ratings (four conditions, each with fifth and twenty-fifth round). Figure 3.8 visualizes the data from the evaluation. Results from a paired-samples t-test showed that children were considered to be significantly less engaged in the twenty-fifth round ($M = 4.38, SD = .84$) than in the fifth round ($M = 5.21, SD = .64$), $t(71) = -12.09, p < .001$. Furthermore, a two-way ANOVA with tutoring strategy (adaptive versus non-adaptive) and gesture use (gestures versus no gestures) as factors showed no significant effect for the use of gestures, $F(1, 68) = 1.36, p = .25, \eta_p^2 = .02$, but there was a significant effect for tutoring strategy, $F(1, 68) = 86.26, p < .001, \eta_p^2 = .559$. The drop in engagement between round five and round twenty-five was less when an adaptive strategy was applied ($M = -.40, SD = .35$) than when words were randomly presented ($M = -1.27, SD = .44$). There was no interaction effect between gestures and tutoring strategies, $F(1, 68) = .01, p = .93, \eta_p^2 = .00$. The same analysis was conducted with the average engagement level of the fifth and twenty-fifth rounds combined, to get an idea of the overall engagement throughout the entire training session in different conditions. In this case the overall level of

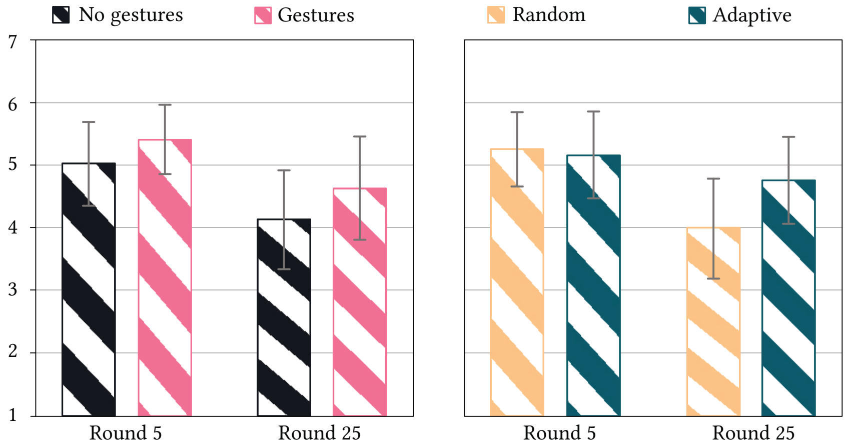


Figure 3.8: Rated engagement levels early and late in the training interaction for the gesture versus no gesture conditions (left) and the adaptive versus random conditions (right), ranging from 1. Completely disengaged to 7. Completely engaged. The error bars are +/- 1 SD.

engagement was significantly higher in the gesture condition ($M = 5.02, SD = .63$) than in the condition without gestures ($M = 4.57, SD = .68$), $F(1, 68) = 8.75, p = .004$, $\eta_p^2 = .114$. There was also a significantly higher engagement when an adaptive strategy was used ($M = 4.97, SD = .67$) as opposed to a random tutoring strategy ($M = 4.63, SD = .67$), $F(1, 68) = 5.10, p = .03, \eta_p^2 = .07$. No interaction effect between the two factors was found, $F(1, 68) = .08, p = .78$, $\eta_p^2 = .001$.

### 3.4.2 Exploration of the effect of age

In the studies described in Chapters 4 and 6, we found that the older children had significantly higher learning outcomes than the younger children, but only when the robot used iconic gestures. Because these studies all had participants of a similar age to the current study (4–6 years old), we explored whether a similar effect of age occured here as well. Chapter 4 does not report on engagement, but Chapter 6 further shows a significant effect of age on engagement with the task, where older children were on average more engaged with the task compared to younger children. There was no significant effect of age on (social) engagement with the robot. In the following section, we investigate the potential effect of age on children's levels of engagement in the current study, although a different method of measuring engagement was used compared to Chapter 6. In the current study, we do not distinguish between engagement with the task and with the robot, but rather use one overall measure. In addition, in the current study engagement was rated by means of an online study, while in Chapter 6 a coding scheme was used to annotate engagement levels. Note that this section was not part of the original published article, and was added as part of this thesis chapter.

#### Effect of age on learning outcomes

Figure 3.9 shows a linear fit to children's age on the x-axis, and their difference scores on the immediate (left) and delayed (right) post-tests on the y-axis, for the experimental conditions with and without gestures. For the immediate post-test, the average difference scores of older children in the conditions with iconic gestures is lower than the scores of younger children. For the delayed post-test, the older children in the conditions with iconic gestures on average learned more words than the younger children. For the conditions without gestures, the learning outcomes appear to be independent of age, on both the immediate and the delayed post-tests.

To ensure consistency with Chapter 6, we conducted a repeated measures ANOVA

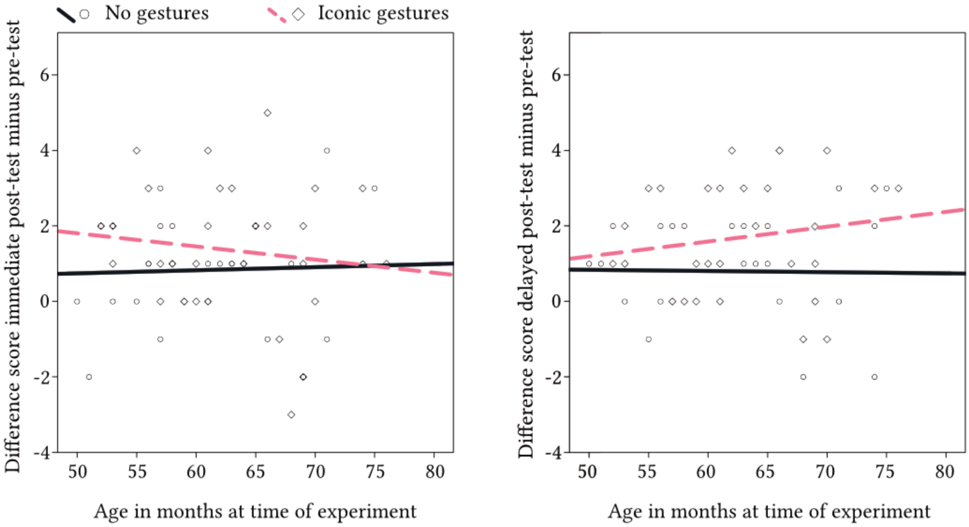


Figure 3.9: Linear fit to the difference scores on the immediate (left) and delayed (right) post-tests compared to the pre-test for the conditions with and without iconic gestures, relative to children's age.

to measure learning outcomes, instead of the two-way ANOVA with difference scores presented in Section 3.4. This repeated measures ANOVA includes scores on the pre-test, immediate post-test, and delayed post-test as dependent variable (with three different measurement points as 'time' within-subjects factor), and the robot's use of iconic gestures (yes or no) as independent variables. The results show a main effect of time, $F(2, 118) = 27.50, p < .001, \eta_p^2 = .32$, indicating that children learned from the tutoring interaction, consistent with Section 3.4. Furthermore, there was a significant interaction effect of time and the robot's use of iconic gestures, $F(2, 118) = 3.30, p = .05, \eta_p^2 = .05$, which is also consistent with Section 3.4.

We then ran the same analysis, now with age as covariate. The results no longer show a main effect of time, $F(2, 116) = 0.58, p = .56, \eta_p^2 = .01$. The interaction effect of time and the robot's use of iconic gestures was also no longer significant, $F(2, 116) = 2.96, p = .056, \eta_p^2 = .05$. There was no main effect of age, $F(1, 58) = 1.01, p = .32, \eta_p^2 = .017$. The interaction of time and age was also not significant, $F(2, 116) = 0.48, p = .62, \eta_p^2 = .008$.

After splitting the data based on whether the robot performed iconic gestures (yes/no), and then conducting the same repeated measures ANOVA, there was no main effect of age for the conditions with iconic gestures, $F(1, 28) = .15, p =$

$.70, \eta_p^2 = .005$, nor without iconic gestures $F(1, 29) = .96, p = .34, \eta_p^2 = .032$. There was also no significant interaction effect of age and time within either subsets, $F(2, 58) = .06, p = .95, \eta_p^2 = .002$ for the conditions without gestures, and $F(1.71, 47.89) = 1.69, p = .20, \eta_p^2 = .06$ for the conditions with gestures (with Huynh-Feldt correction because the assumption of sphericity was violated). This is contrary to Chapter 6, where a significant interaction effect of time and age was found for the experimental conditions involving robot-performed iconic gestures.

### Effect of age on engagement

Again for consistency with Chapter 6, we conducted a repeated measures ANOVA with the average level of engagement per child on a 7-point scale (1. completely disengaged – 7. completely engaged), as rated by adult participants in our evaluation study, as dependent variable (with round 5 and round 25 measurements as 'time' within-subjects factor), and the robot's use of iconic gestures (yes/no) as independent variable. This shows a significant main effect of time, $F(1, 58) = 17.88, p < .001, \eta_p^2 = .24$, where children on average were less engaged in round 25 ($M = 4.37, SD = 1.31$) than in round 5 ($M = 5.22, SD = .79$). In addition, there was a significant main effect of the robot's use of iconic gestures, $F(1, 58) = 6.22, p = .02, \eta_p^2 = .097$, where children in the conditions with iconic gestures were on average more engaged throughout the interaction ($M = 5.03, SD = 0.71$) compared to children in the conditions without gestures ($M = 4.56, SD = 0.75$). The interaction effect of time and the robot's use of iconic gestures was not significant, $F(1, 58) = 0.12, p = .73, \eta_p^2 = .002$. All of these findings are consistent with the paired samples t-test and two-way ANOVA presented in Section 3.4.

After adding age as a covariate and conducting the same repeated measures ANOVA with engagement in round 5 and 25 as dependent variable, and the robot's use of iconic gestures as independent variable, the main effect of time was no longer significant, $F(1, 57) = 2.42, p = .13, \eta_p^2 = .041$. The main effect of the robot's use of iconic gestures was significant, $F(1, 57) = 6.20, p = .02, \eta_p^2 = .098$. The interaction effect of time and the robot's use of iconic gestures was not significant, $F(1, 57) = .08, p = .78, \eta_p^2 = .001$. There was no significant main effect of age, $F(1, 57) = 0.13, p = .72, \eta_p^2 = .002$, nor an interaction effect of time and age, $F(1, 57) = 1.24, p = .27, \eta_p^2 = .021$. This indicates that there is no effect of age on the observed levels of engagement.

## 3.5 Discussion

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

When gestures were performed by the robot during training, there was a higher retention of newly acquired words after at least one week. This aligns with similar effects that were shown previously in the context of math with a human tutor (Cook et al., 2008) and indicates that these indeed carry over to a robot; a compelling finding that warrants future research into the intricacies of gesture use by humanoid robots. As mentioned by Hostetter (2011) with respect to human-human communication, it appears that gestures retain their positive effects on communication when they are scripted rather than being produced spontaneously. In this work, only iconic gestures are used that clearly relate to the concept they describe. Future work could investigate whether a similar contribution to learning gain is found when non-iconic gestures are used. Furthermore, the target words used in this experiment were chosen specifically such that matching gestures could be designed for the robot. It would be interesting to explore how well a broader range of gestures, describing various abstract and concrete concepts, could be performed by a robot as opposed to a human interlocutor. Finally, asking children to actually re-enact the gestures (e.g., as in de Nooijer et al., 2013; Tellier, 2008), or to come up with their own gestures, might further increase the potential utility of gestures in learning due to the embodiment effect (Dijkstra & Post, 2015).

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

We found it valuable to include the measure of children's engagement during the interaction. A higher level of engagement indicates increased motivation and willingness to learn (Blumenfeld et al., 2005). Although students might succeed in simple word learning with limited engagement and the use of a low-level learning strategy, increased engagement could stimulate them to go beyond simple memorization and relate these new words to prior knowledge. Furthermore, engagement can serve as a measure of how well the learning activities are tailored to the child's abilities — constantly presenting tasks that are either too hard or too easy could have a detrimental effect on engagement. The results of our evaluation show that indeed the adaptive system appears to match the learning activities to each child's needs by providing a realistic yet challenging task, resulting in a reduced decline in engagement toward the end of the interaction. Gestures contribute to a higher overall engagement, which could be explained by the fact that the robot appears more active and playful in this condition, thereby stimulating the child to remain engaged.

## 3.6 Conclusion

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

***

*Our first study into robot-performed iconic gestures to support second language learning, presented in this chapter, showed promising results: children showed higher levels of engagement and retained more English words if the robot used iconic gestures to support its tutoring efforts. However, the English animal names in this study had expressive, clear gestures, and therefore we were left wondering whether these observed effects would remain if children were learning more abstract words, relating to a broader range of semantic categories. This is addressed in Chapter 4. Furthermore, the study presented in Chapter 4 contains more diverse activities for the child and robot to engage in (e.g., repeating words), and it consists of multiple sessions to see whether a robot tutor can support second language learning on the longer term.*

4

# Chapter 4

# Designing and Evaluating Robot-Performed Iconic Gestures

*This chapter is based on:* de Wit, J., Willemsen, B., de Haas, M., van den Berghe, R., Leseman, P., Oudgenoeg-Paz, O., Verhagen, J., Vogt, P. & Krahmer, E. Designing and evaluating iconic gestures for child-robot second language learning. Submitted for journal publication. *Open practices:* The materials are available at https://github.com/l2tor/. The study and analysis plan were preregistered on AsPredicted: https://aspredicted.org/6k93k.pdf.

## Abstract

In this chapter, we examine the process of designing robot-performed iconic hand gestures in the context of a long-term study into second language tutoring with children of approximately five years old. We explore several factors that may influence their efficacy in supporting second language tutoring: the age of participating children, differences between gestures for various semantic categories — e.g., measurement words, such as small, versus counting words, such as five — the quality (comprehensibility) of the robot's gestures, and spontaneous reenactment or imitation of the gestures. Age was found to affect children's learning outcomes, with older children benefiting more from the robot's iconic gestures than younger children, particularly for measurement words. We found no conclusive evidence that the quality of the gestures or spontaneous reenactment of said gestures affected their ability to improve learning outcomes. We further propose several improvements to the process of designing and implementing a robot's iconic gesture repertoire.

## 4.1 Introduction

There is an increasing interest in the use of robots for educational purposes (Belpaeme et al., 2018; Mubin et al., 2013; Toh et al., 2016). They can be used as a subject of learning, for example by building and programming robots together with students to teach them about robotics, artificial intelligence, or computer programming. Alternatively, *social* robots can take on the role of tutors by presenting educational content and engaging in teaching activities in a multitude of domains (Mubin et al., 2013), including language learning — which is the focus of the current study. One of the main motivations that drive the use of technology, and social robots specifically, in education is the need to accommodate individual children's diverse needs while at the same time the average number of students per teacher is increasing (Blatchford & Russell, 2020). By working alongside teachers (and certainly not replacing them), robots can present a cost-effective way of expanding and personalizing the content that can be offered to learners. On top of the functional role of presenting educational content — which can also be done with other tools such as tablets — robots are arguably able to provide additional social support, for example by providing (non-verbal) feedback and giving empathic responses rather than focusing merely on knowledge transfer, which has been shown to enhance the learner's engagement, as well as learning outcomes (Saerbeck et al., 2010).

An important part of a robot's perceived social intelligence (Fong et al., 2003) is its ability to use non-verbal communication such as gestures. Pointing (*deictic*) gestures, for instance, can be used to guide the attention of the learner toward the educational content, by referring to relevant objects (Sauppé & Mutlu, 2014). *Iconic* gestures, which are closely related in shape or motion to the concept being described (McNeill, 1992), can be used to ground new knowledge in familiar concepts or actions from the real world (Barsalou, 2008). For example, a *ball* can be depicted by molding a sphere with one's hands (shape), or by kicking an imaginary ball (motion). One particular domain that appears to benefit from gestures is (second) language learning (Hald et al., 2016; Rohlfing, 2019), a domain that has recently also gained considerable attention from research into educational robots (see e.g., Kanero et al., 2018b; van den Berghe et al., 2019, for overviews of existing work). In second language learning, gestures can be used as a bridge between unknown words in the second language and existing knowledge of concepts or experiences (Hald et al., 2016). In other words, gestures can be used to link the learner's non-linguistic (e.g., motor, visual) knowledge of a concept to the linguistic form of said concept.

However, robot-performed gestures may have to look different from what we are used to, as current commercially available robots are more limited in their motor degrees of freedom than humans. For example, most commonly used robots are not able to move individual fingers, making it hard to perform finger-counting or detailed hand gestures (Vogt et al., 2017a). This means that in many cases it is not possible to directly copy the way humans perform a gesture onto a robot, potentially resulting in a loss of information which reduces the communicative ability of the gesture. This raises the question whether robot-performed gestures are able to provide the same beneficial effects to learning that we see with human-performed gestures (e.g., Hostetter, 2011; Roth, 2001).

In a previous study we investigated whether a NAO humanoid robot could support its second language tutoring efforts with iconic gestures. We found that children of four to six years old retained more words over time, and were more engaged during the interaction if a robot used iconic gestures when introducing words in the second language, as opposed to a robot that did not use such gestures (Chapter 3). In a follow-up to this previous work, which will provide the basis for the current chapter, we have made several adjustments to the set-up of the study: instead of the highly iconic animal names that were taught in the previous study, the follow-up included concepts for which it is more challenging to come up with gestures with a high degree of iconicity, such as prepositions (*next to*) and comparatives (*most*). Additionally, the follow-up study consisted of seven sessions with the robot, instead of the single session in the first study. The follow-up study was conducted with children of a similar age group to the previous study, and it included a larger sample. In this case, however, we found no effect of the robot's use of iconic gestures on children's learning outcomes (Vogt et al., 2019).

These mixed findings across the two studies, combined with the overall positive results found in literature on both human-performed and robot-performed gestures in supporting language learning (e.g., Hald et al., 2016; van Dijk et al., 2013), show us that it is important to carefully consider the design and implementation of the robot's gestures, and to investigate any contextual factors that may have prevented children from benefiting from them in our second study.

Based on existing studies into human gesturing, we have identified four factors that may influence the effectiveness of robot-performed gestures in the context of education. First, iconic gestures only appear to contribute to learning if their meaning is clear, and congruent with what is conveyed via speech (Kelly et al., 2009; Macedonia et al., 2011). It is therefore important that the gestures are designed in such a way that they are comprehensible for the learner. Second, the ability to interpret the meaning of iconic gestures develops during a child's early years (Novack et al., 2015; Stanfield et al., 2014). Based on the literature, the children in our study were generally at an age (5–6 years old) where they should be able to interpret the gestures. However, the fact that they were performed by a robot, with certain physical limitations and a different morphology from humans, might have negatively affected this ability. Age could therefore have played a role in the effectiveness of the robot's gestures. Third, various studies have shown indications that gestures may have a greater contribution for teaching the linguistic forms of certain types of concepts (e.g., motor events such as *running*), compared to others (de Nooijer et al., 2013; Hostetter, 2011). Finally, research with human-performed gestures in the context of language learning suggests that reenactment or imitation of the teacher's gestures by the learner could further strengthen their contribution to the learning process (Repetto et al., 2017; Tellier, 2005). Based on these previous studies, we pose the following research question:

**(RQ)** To what extent do the comprehensibility of the robot's gestures, the age of participating children, different semantic categories, and gesture reenactment influence the successful application of robot-performed iconic gestures in second language tutoring for children?

In the current chapter, we build upon our previous study (Vogt et al., 2019). This is done, firstly, by thoroughly reflecting upon and evaluating the design of the robot's gestures, in order to find ways to improve the gesture design process. Secondly, we provide additional analyses of the data that were previously collected, focusing specifically on the four aforementioned contextual factors: comprehensibility, age, concept-based differences, and reenactment. Our aim with this work is to present concrete guidelines for the design and implementation of iconic gestures for social robots, in order to optimally make use of the beneficial effects that the robot's gestures could have on (second language) learning. In the following sections we provide an overview of existing research in the field of robots for education and gestures, and we cover previous work that investigated gestures performed by robots, particularly focusing on studies in education. We then introduce the set-up of the experimental study that was conducted in order to investigate the effects of a robot's use of iconic gestures to support second language learning, from which the data are used in the current study, and describe in detail the process of designing the robot's iconic gestures. Finally, we present and then discuss the results of our evaluation of the comprehensibility of the robot-performed iconic gestures, as well as the role of age, item-based differences, and reenactment.

## 4.2 Background

### 4.2.1 Social robots in education

The potential use of social, humanoid robots in education has become a recent focus of attention in research and in society. Next to their functional goal of presenting educational content, robots are also able to fulfill a social role that is conducive to learning, because people tend to assign human-like characteristics to them (Duffy, 2003), and therefore want to communicate with them in a human-like way (Bartneck & Forlizzi, 2004). This enables robots to teach meta-cognitive skills such as thinking aloud that can further support learning (Ramachandran et al., 2018). A socially intelligent robot (Fong et al., 2003) is able to observe the emotions of others and adjust its behavior accordingly (Gordon et al., 2016; Szafir & Mutlu, 2012), and it can also display emotions of its own, thus showing a certain personality or character (Breazeal, 2004; Robert et al., 2020). Furthermore, it is able to engage in a dialogue with human interlocutors using natural language, and support its communication with non-verbal behavior such as gaze and gestures (Anzalone et al., 2015; Scassellati, 2002). Its socially intelligent behavior enables the robot to build rapport, which in turn elicits more social behavior, such as constructive help-seeking (Howley et al., 2014), from the learner as well. The bond between robot and learner can be further strengthened by personalizing the interactions, for example by addressing learners by their names and engaging in small talk by asking them about their interests. This can stimulate others to open up and engage more with the robot (Henkemans et al., 2013). However, research by Kennedy et al. (2015) shows that caution is advised when designing educational human-robot interactions, as it is also possible for a robot to become *too* social, which could have a detrimental effect on learning.

Compared to virtual agents that can offer similar advantages in education, robots additionally have a physical presence in the context of the learner, which is suggested to stimulate social behavior and result in greater learning gains (Belpaeme et al., 2018). A robot that is physically present is also generally rated more positively, and regarded as more persuasive than a telepresent robot that is displayed on a screen or a virtual agent (Li, 2015). Furthermore, people are more likely to comply with tasks that can be seen as unusual (e.g., putting books in the trash), which rely heavily on trust (Bainbridge et al., 2011), when these tasks are presented by a physically present robot instead of a virtual agent. Depending on the educational domain in which the robot is active, its ability to move within and interact with the physical world could be used to support its teaching activities (Özgür et al., 2017), for example by providing realistic feedback on tasks that require manipulations in the physical world, and it allows the robot to perform classroom management (Kanda et al., 2012).

One particular educational domain in which robots are commonly deployed is language learning. Because robots are seen as socially present entities, it is possible to create an immersive, natural context where learners can engage in conversations with the robot in order to facilitate language learning by immediately applying newly acquired skills in practice while receiving feedback (S. Lee et al., 2011). Chang et al. (2010) further highlight the robot's ability to tirelessly repeat content, and the potential use of body language to support language learning. A study by Alemi et al. (2015) reports that children felt less anxious, were more motivated, and reported higher levels of enjoyment when training second language vocabulary with a robot compared to when no robot was present.

Previous research by Han et al. (2008) investigated the difference between children learning a second language from a robot, web-based instruction, and a book with audiotape in the context of their homes. Content was kept similar by taking the design for the web-based instruction and turning it into static imagery for the book, and by displaying it on the robot's embedded tablet screen. They found that children were more interested and focused, and performed better when a robot was used. However, Westlund et al. (2015) compared language learning from a robot, tablet, and human teacher, and did not find any differences in terms of learning outcomes, although children did indicate that they preferred learning from the robot over the tablet and human teacher.

To summarize, existing research shows promising results regarding the use of social robots in education. Their physical embodiment and presence in the context of learning set robots apart from other educational tools, such as tablet devices. Gestures could form an important way to make use of the robot's physical presence.

### 4.2.2 Gestures in education

Gestures are generally defined as "visible actions" portrayed with our bodies (Kendon, 2004). The use of gestures plays an important role in our communication with others, for example by guiding the attention of listeners, and by making it easier for them to understand information that is communicated verbally (Hostetter, 2011). In communication, we use different types of gestures, including rhythmic *beat gestures* to emphasize certain parts of our speech, *deictic gestures* such as pointing to direct attention toward a specific entity, and *representational* or *iconic gestures* in which the hands or body are used to depict a particular action, object or concept that may not be physically present (McNeill, 1992). The concept that an iconic gesture refers to is represented in some way by the motion itself, for example by pretending to brush our teeth when trying to describe a *toothbrush*, or by molding the shape of an imaginary *ball* in the air. In the current study the robot employed occasional deictic gestures to guide attention, but we focus mainly on investigating the use of iconic gestures.

Gestures, and iconic gestures in particular, are often used spontaneously and together with speech, although silent gesture or pantomime that act as a substitute for speech occur as well (McNeill, 1992). The use of gestures is an important tool in educational settings (Kelly et al., 2008), where it can be considered a form of scaffolding that helps the learner understand the materials, which is particularly useful when concepts are complex or newly introduced (Alibali & Nathan, 2007). Additionally, teachers are able to hold the students' attention for longer periods of time when they use gestures to support their teaching (Valenzeno et al., 2003). Specifically in second language learning, gestures can serve as a bridge between a concept that is familiar to someone in their native language, or L1, and its still unknown translation in the second language, L2, by grounding the new L2 word in existing knowledge of actions or objects (Barsalou, 2008).

#### Meaningful, comprehensible gestures

Several studies have examined the added value of iconic gestures for (second) language learning (see e.g., Hald et al., 2016; Rohlfing, 2019, for a review). For example, Kelly et al. (2009) compared between L2 word learning without support from gestures, without gestures but with repeated speech, with congruent gestures, or with incongruent gestures (which were the same gestures as in the congruent condition, but produced with other words than to which they belonged). Participants who received support from congruent gestures learned most words, followed by the group that received repeated speech input, the group without any additional cues, and lastly the group that received incongruent gestures. Macedonia et al. (2011) conducted a study in which they compared between the use of iconic gestures and meaningless gestures to support learning of an artificial language, and they found that using iconic gestures resulted in better learning outcomes than when meaningless gestures were used. Both studies show that the role of iconic gestures goes beyond merely drawing attention to the speaker, and that it is relevant to design gestures in such a way that they communicate the right meaning. Based on this, we pose the following subquestion to guide our research:

**(Q1)** How does the comprehensibility of the robot's iconic gestures affect their contribution to learning?

### Age of the learner

We learn to interpret iconic gestures at a relatively young age. Novack et al. (2015) compared between two- and three-year-old children, and found that two-year-olds could already take advantage of iconic gestures in the context of learning how to use new toys, although not as much as three-year-olds. Another study by Stanfield et al. (2014) found that children start to understand non-redundant iconic gestures (e.g., a combination of "read" in speech with an iconic gestures for *book*) by age three, and that this skill continues to develop as they grow older. Existing research highlights a number of additional factors that may influence the effects of iconic gestures on communication and learning. For example, children with weaker L1 skills generally benefit more from gestures than people that have stronger L1 skills (Rowe et al., 2013). Children were found to especially find support in gestures when the spoken part of the message was complex (McNeil et al., 2000), potentially also due to their still developing language skills. These individual differences, particularly at a younger age as our ability to interpret gestures is still developing, lead us to the second subquestion:

**(Q2)** What is the role of age in the effects of the robot's iconic gestures on learning?

### Concept-based differences

It is further suggested that the positive effects of iconic gestures are stronger when they describe spatial concepts (e.g., spatial relations such as *under*) or motor events (e.g., actions such as *running*) than when the concepts are more abstract, such as colors, where the link between the motions and the referent is less clear (Hostetter, 2011). However, a study by Repetto et al. (2017), in which young adults were taught a number of abstract words (e.g., *boredom* and *alternative*) in an artificial language, still showed that participants remembered more words when they were presented to them in combination with gestures, than when the words were presented with pictures or with no additional cues. Research further suggests that verbs are especially challenging for children to learn, because children have difficulty generalizing from the particular objects or context with which they were originally taught. Because gestures do not involve interactions with real physical objects, they support the acquisition of generalizable verb knowledge better than actually performing the action on a specific tangible object (Wakefield et al., 2018). In summary, research on potential differences in the effectiveness of gestures based on concept or word types is scarce, but provides a first indication that such differences do exist. As a result, we pose the following subquestion for the current research:

**(Q3)** Are there (item-based) differences in the contribution of gestures in supporting learning, depending on the types of concepts that are depicted?

### Gesture reenactment

One important aspect of the study by Repetto et al. (2017), that might support learning by means of gestures, is that participants were asked to reenact or imitate the movements after observing them on screen, rather than merely observing them. In a study by Cook et al. (2008) children of eight to ten years old were asked to mimic the instructor's behavior when solving mathematical problems, which led to better long-term retention of the instructions compared to children that did not perform gestures themselves. Tellier (2005) found similar effects, first in the context of L1 vocabulary learning, where 42 children (five to six years old) were split into three groups: One group was asked to repeat the words, the second also repeated the words and observed matching gestures, while the third group repeated words and imitated the gestures. The group that mimicked gestures performed significantly better in a short-term recall test than both other groups.

In a follow-up study (Tellier, 2008), twenty children within the same age group as the previous study learned L2 vocabulary over the course of multiple sessions. They either received pictures of the concepts that the words related to as support, or video recordings of people performing gestures for these concepts. If they were shown gestures, the children were asked to imitate them. The group of children who encoded the words using gestures performed better on the assessments, particularly on tests of their active knowledge (production, rather than recognition of the L2 words), than the group who observed pictures (Tellier, 2008). In a study by de Nooijer et al. (2013), children of nine to eleven years old learned L1 verbs and were divided into four groups. One group only observed matching gestures while training the words, while the other three groups imitated the gestures, either 2) during training, 3) while trying to recall the verbs on the post-test, or 4) in both situations. The results of this study indicated that imitation was only helpful for the object-manipulation verbs that were present in the study, and not for the locomotion or abstract verbs.

These findings regarding the potential benefits of enacting in order to memorize concepts align with the notion of embodied cognition, and the language-action connection (Glenberg & Gallese, 2012). Although, to our knowledge, there is no existing research that draws a direct comparison between observing and reenacting iconic gestures in the context of L2 learning, based on findings in other educational domains and L1 learning we expect that children who (spontaneously) reenacted gestures in the current study may have benefited more from them than those who did not reenact, therefore we pose the following subquestion:

**(Q4)** Does reenactment (mimicry, imitation) of the robot's iconic gestures by the learners improve learning outcomes?

To summarize, iconic gestures have proven to be valuable tools to support education, particularly in the domain of second language learning. Their contribution to learning appears to be dependent on several factors, including the characteristics of the learner, the materials that are being taught, and whether the gestures are merely observed or also imitated. We aim to investigate whether these same factors play a role in human-robot interaction.

### 4.2.3 Related work on robots and gestures

Because robots are generally more limited in their motor degrees of freedom, their gesturing capabilities are not as extensive as that of humans, or modern virtual agents that are driven by motion capture recordings. This raises the question whether robots are expressive enough to be able to leverage the aforementioned benefits that gestures provide in human-human communication, specifically in educational contexts. Bremner and Leonards (2016) compared between co-speech iconic gestures produced by a human, and the same gestures copied to a robot using motion capture techniques. They found that for most gestures the participants in their study were able to identify the meaning in a multiple choice task equally well, regardless of whether they were performed by a human or a tele-operated robot.

Not only do the robot's gestures appear to support its communicative efforts, robots that add a non-verbal component to their speech output are also perceived differently from those that do not. A study by Salem et al. (2013a) found that a robot that used gestures was perceived as more human-like and likeable than one that did not gesture, even more so when the robot made errors by performing motions that were incongruent with its speech, although at the cost of task performance. Gestures can also be used to give a certain personality or emotional state to the robot, which in turn could lead to richer, more personal interactions and to further improve people's attitude toward the robot (Aly & Tapus, 2013; Craenen et al., 2018). Furthermore, several studies have reported higher levels of engagement when robots use gestures, compared to when they are static or perform random movements (Bremner et al., 2011; Chapter 3). In a review by Li (2015), the results from several studies indicate that people's attitude tends to be more positive toward a physically present robot compared to one that is telepresent (i.e., displayed on a screen) and to virtual agents, but only when it is using gestures — the opposite effect was found when the robot did not use gestures. This is an indication that one of the main advantages of a robot that is physically present over virtual alternatives is that it is able to move and communicate in the real world context.

Ahmad et al. (2016b) conducted an interview study with primary and high school teachers. The teachers agreed that social robots could be useful for language learning, and they stressed the importance of gestures in language education (with and without robots). Empirical research specifically into the effects of a robot's use of iconic gestures in the context of (second) language learning is however still scarce. In a study from the related field of information retention, van Dijk et al. (2013) showed in a single session with adult participants that the use of iconic gestures by a robot increased retention, particularly of verbs, measured using a recall task. Similar results on information retention were found in the context of storytelling (Bremner et al., 2011; Huang & Mutlu, 2013). Another study involving storytelling by a robot further suggests that exaggerated gestures, which are perceived as more cartoon-like, lead to increased memorization of the story compared to "normal" (unexaggerated) motion, and the robot was perceived as more engaging and entertaining when exaggerating its movements (Gielniak & Thomaz, 2012).

In previous work with four- to six-year-old children, we have shown that the robot's use of iconic gestures while presenting words (animal names) in a second language aided the recall of these words approximately one week after training, and resulted in an overall higher level of engagement of the child while learning with the robot (Chapter 3). Although these gestures were intentionally chosen and designed to have a high degree of iconicity, the results of this study do serve as a first indication that the benefits of iconic gestures that we see in human-human tutoring situations could apply to robot-performed gestures as well. After this initial exploration, we conducted a large-scale study to further investigate the potential application of social robots in second language tutoring. In Vogt et al. (2019), we have concisely described the learning effects in the different conditions (briefly summarized in the next section), which provides the basis for the current chapter. In this chapter, we present an in-depth analysis of the design and the effects of the robot's use of iconic gestures, which was not part of Vogt et al. (2019).

### 4.2.4 Large-scale study

We conducted a study at nine different primary schools throughout the Netherlands, in which children of approximately five years old ($M = 5$ years, 8 months; $SD = 5$ months) interacted with an intelligent tutoring system (ITS), consisting of a tablet device on which educational content was shown, and a robot that engaged in learning activities with the children. The study included seven sessions, where new L2 vocabulary was introduced in the first six, while the seventh session served as a recap of the previously taught words. Our aim was to investigate: (1) whether the intelligent tutoring system is effective at teaching children L2 vocabulary; (2) whether the robot's physical presence contributes to learning outcomes; and (3) whether robot-performed iconic gestures result in greater learning outcomes, compared to a robot that does not use iconic gestures. In order to study these effects, we assigned the children to one of the following conditions:

1. **Control (no treatment)**, where children had an interaction with the robot once a week (for a total of three interactions), which did not involve any educational content related to second language vocabulary.

2. **Tablet only**, where children interacted only with the tablet. The robot was hidden from view, with its speech output routed through the tablet's speakers.

3. **Tablet + robot without iconic gestures**, where children interacted with the tablet and the robot, and the robot would use deictic (pointing, tablet manipulation) gestures to guide the child's attention.

4. **Tablet + robot with iconic gestures**, where children interacted with the tablet and the robot, and the robot would use both deictic gestures to guide the child's attention and manipulate objects on the tablet, as well as a matching iconic gesture whenever it pronounced one of the target words in the second language.

A total of 194 children, 97 boys and 97 girls, participated and met the inclusion criteria (e.g., scoring a maximum of 17 out of 34 words correct on the English translation pre-test). They were pseudo-randomly assigned to the experimental conditions, with a balance in age and gender, resulting in 32 participants in the control condition (1), and 54 participants in each of the three experimental conditions (2–4). The children's legal guardians gave informed consent, and the study was approved by our institutions' research ethics committees. The study and analysis plan were preregistered on AsPredicted[1].

The results, which are presented in detail in Vogt et al. (2019), showed that children in the three experimental conditions scored significantly higher on translation as well as comprehension tasks, than those in the control condition (all $p$-values $< .01$). This means that the tutoring interaction was effective. However, contrary to our expectations, no significant differences were found between the three experimental conditions of tablet only, tablet + robot without iconic gestures, and tablet + robot with iconic gestures. In other words, there was no observed effect of the robot's physical presence and use of deictic gestures, nor of its use of iconic gestures, on the students' learning outcomes. For the remainder of this chapter, we will focus our attention on the robot's use of iconic gestures, to get a better understanding of the role of these gestures in the child-robot interactions.

In the following section, we first describe the design of the intelligent tutoring system as a whole. This is important, because the iconic gestures were included as part of this tutoring interaction and were not used in isolation, therefore the nature of this interaction (and how it is different from other studies) could potentially have influenced the effectiveness of the robot's iconic gestures. We then introduce the process of designing the gestures, and what the resulting gestures looked like. The

[1] https://aspredicted.org/6k93k.pdf

measurement instruments are then presented, as these are needed to interpret the analyses that follow. Finally, we present the results of our analyses, and conclude with a discussion of our findings and recommendations for the design and implementation of robot-performed iconic gestures.

## 4.3 Interaction and gesture design

### 4.3.1 Design of the tutoring interaction

The intelligent tutoring system (ITS) consisted of a Softbank Robotics NAO V5 robot, combined with a Microsoft Surface Pro 4 tablet through which the child engaged in the learning interaction. The robot was placed in a crouching position at a 90-degree angle relative to the child. This helped to position the robot as a peer rather than a teacher, which has been shown to result in increased task engagement and performance (Zaga et al., 2015). In addition, this made it easier for the learner to take on the robot's perspective, thereby avoiding confusion for gestures such as *left*, which would be harder to interpret if the robot would be sitting directly across from the learner. Figure 4.1 shows the general positioning of the robot and tablet. This positioning was kept as consistently as possible between different schools. One camera was placed facing the child, with a second camera to the side and behind the child, so that the interactions with the tablet could also be recorded. To make the robot seem more life-like, we enabled "breathing mode" which caused its arms to move around slightly, giving the illusion that the robot was actively breathing. It also blinked its eyes every few seconds, and was tracking the child's face to establish eye contact.

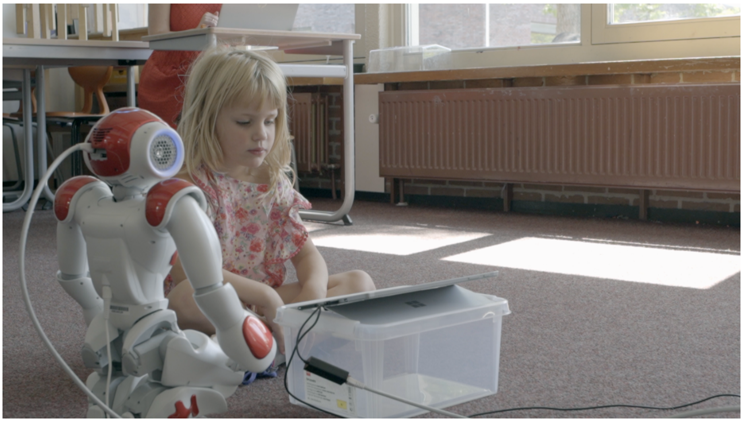


Figure 4.1: Positioning of the tablet and the robot during the experiment.

The content of the study comprised seven lessons in total. The first six lessons each took place in a different virtual environment, such as a forest or a playground, where the native Dutch-speaking child was introduced to 5–6 new words in the second language (English) during each lesson (see Figure 4.2 for an example). We opted for the use of virtual environments and objects instead of physical ones because automatic perception and manipulation of real objects in a dynamic physical context would have been challenging to implement. Virtual objects have also been shown to be equally effective in supporting math and L1 teaching (Klahr et al., 2007; Singer & Gerrits, 2015). Moreover, in a preliminary study comparing the effects of physical versus virtual objects on L2 learning, we did not find differences in learning outcomes (Vlaar et al., 2017).

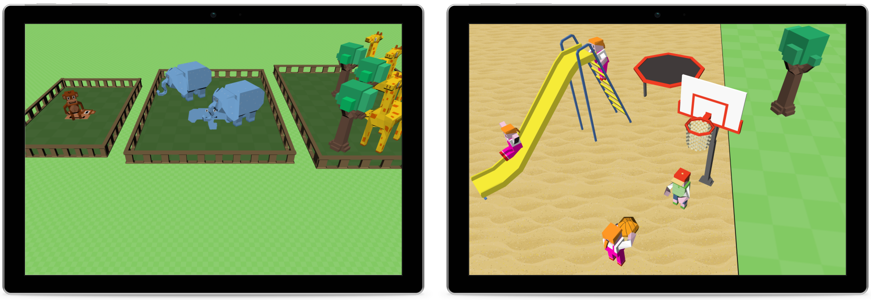


Figure 4.2: Examples of the virtual environments shown on the tablet. Left: lesson one in the zoo, where animals have been brought back to their cages. Right: lesson six in the playground, which was first 'built' by placing equipment, and now children started playing in the area.

In the first three lessons, the target words belonged to the number domain, including concepts such as counting words (*one, two, three, four, five*), mathematical operations (*add, take away*) and comparisons (*more, most*). Lessons four, five and six focused on spatial relations and verbs, which contained words such as *above, next to, walking* and *sliding*. These words were selected based on a survey of existing educational curricula, word frequency and age of acquisition lists[2] to ensure that children were familiar with the concepts in their native language. The final seventh lesson did not introduce any new target words, but instead recapitulated all 34 target words from the previous six lessons. Table 4.1 shows a list of all the English words that were included in the study, as well as the virtual environment in which they

[2] https://web.archive.org/web/20210415022714/http://www.l2tor.eu/effe/wp-content/uploads/2015/12/D1.1-Lessons-series-three-domains.pdf

Table 4.1: English words included in the study, per lesson.

| Lesson | Environment | English words |
|---|---|---|
| 1 | Zoo | One, two, three, add, more, most |
| 2 | Bakery | Four, five, take away, fewer, fewest |
| 3 | Zoo | Big, small, heavy, light, high, low |
| 4 | Fruit shop | On, above, below, next to, falling |
| 5 | Forest | In front of, behind, walking, running, jumping, flying |
| 6 | Playground | Left, right, catching, throwing, sliding, climbing |
| 7 | Photo book | Recapitulation of all words |

were presented.

During each lesson, children went through a particular scenario together with the robot, while they completed several different tasks that were presented to them by the robot, such as touching or moving objects on the screen, repeating words after the robot, or performing an action in the real world (such as pretending how to climb). To further position the robot as a peer, the tablet was responsible for actually initiating these tasks, for example by making new objects appear on the scene. The robot would then observe this change on the tablet and suggest the course of action in order to continue, as if the robot and child were learning together, for example by stating that “the monkey has escaped — let’s put it back in its cage!”.

The lessons followed predefined scripts, so that each child experienced the same interaction by performing the tasks in the same order. The scripts were created in such a way that all target words were mentioned at least ten times during the lesson in which they were introduced, and once more in the lesson that followed it. In the condition with iconic gestures, the robot would perform the matching gesture whenever it pronounced a target word in the L2. If the child did not perform any action or if the action was incorrect, the robot would repeat the task up to two times, which resulted in additional exposures to the English words and, in the condition with iconic gestures, the matching gestures. If the task was still unfinished after two reminders, the robot performed the task for the child, for example by moving objects on the screen or by counting down and then repeating the words together with the child, to ensure that the script was always completed. Because of the robot’s imperfect pronunciation, the first mention of each word was by means of a recording from a native English speaker, which was played back through the tablet, generally as a response to the child successfully performing an action such as touching or moving an object on the screen.

During the seventh lesson, which was a recapitulation of the previously learned words, children constructed a photo book that contained six pages, each with a screenshot of the backdrop of one of the previous lessons. The children were then asked to drag stickers containing objects that were present in the original scenes onto the pages, while practicing the related English words. Figure 4.3 shows one of the pages of the photo book, with the stickers not yet placed. While the other lessons all had three-dimensional environments, the recapitulation lesson was two-dimensional. Although all 34 target words had to be covered in this lesson, there were fewer repetitions of these words compared to previous lessons, resulting in a total session length of approximately 15–20 minutes, which was similar to the other six sessions.

The researcher had a control panel running on a laptop, which could be used to start a specific lesson. This was also used to enter the child's name, so that the robot could use it during the interaction, and it provided the researcher with the option to pause the lesson if needed. The robot acted nearly fully autonomously, with the exception of recognizing whether children successfully completed tasks in which they had to repeat words after the robot, or had to enact a certain action, for which the sensing techniques were difficult to implement. For example, the use of automatic speech recognition (ASR) to detect whether children correctly repeated after the robot is not yet reliable enough (Mubin et al., 2012), especially when attempting to recognize young children's speech (Kennedy et al., 2017). For tasks where the child had to repeat a target word, the researcher therefore pressed a button on the control panel when the child spoke for the interaction to continue (a Wizard of Oz approach). For other points during the scenario where we expected a reply from

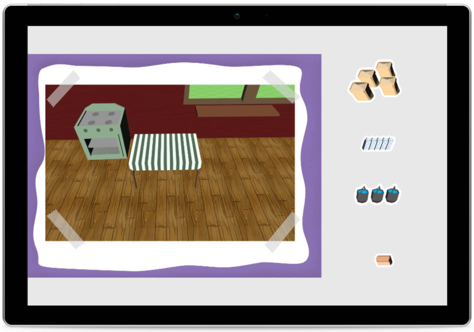


Figure 4.3: The photo book environment used in the seventh (recapitulation) lesson.

children (e.g., during small talk, or in the case of enactment), we implemented pauses to create the illusion that the robot was watching and listening to the children. For an impression of what the interaction between child and robot looked like, we refer to a promotional video that was developed as part of the L2TOR project[3].

## 4.3.2 Design of the robot's gestures

### Deictic gestures

The robot performed three types of deictic gestures during the tutoring interactions with the children (Figure 4.4). The first type was implemented at predefined locations within the script, where the robot would point toward the tablet screen to direct the child's attention to it. This gesture was always the same, so there was no distinction between different parts of the screen — the robot directed its gaze toward the tablet, and pointed in its general direction. The other two types of deictic gestures were used when the robot provided help to the learner after a task was performed incorrectly or not performed at all. If the task was to move an object to a different location, the robot would "swipe" across the screen while at the same time the object would move to its correct target location. A similar motion was implemented to simulate the robot touching an object on the screen. In this case the robot would extend its arm over the tablet and then briefly open and close its hand. At the same time, the corresponding object was highlighted on the screen to simulate the robot's triggering of the object. Both the swiping and touching gestures, just like the pointing gesture, were always the same and were not linked to any exact locations on the tablet. However, this proved to be realistic enough to provide the illusion of the robot performing manipulations within the virtual environment. We also explained to children that this was how the robot controlled the tablet, and this explanation was

[3] https://youtu.be/y8W-2XgdfoI

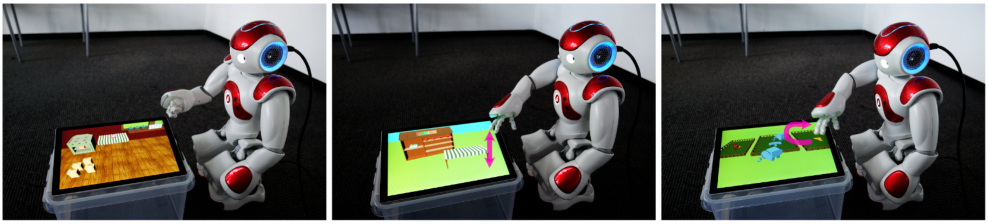


Figure 4.4: The three types of deictic gestures used in the study. Left: pointing (closed hand); middle: pretending to touch the screen (the hand briefly opens and closes); right: pretending to swipe across the screen (open hand).

accepted by them.

### Designing human-like iconic gestures

The iconic gestures for the chosen target words were based on a dataset that was collected using a gesture elicitation procedure (Kanero et al., 2018a). In this elicitation study, three participants, all native speakers of English, were recorded while performing matching gestures for all 34 concepts. Twenty other participants, also native English speakers, were then asked to view these recordings and rate on a scale from 1–7 the comprehensibility of the human-performed versions of the gestures, or the degree to which they matched the words they intended to describe. Because participants in this study were not constrained to the robot’s physical limitations, several gestures contained certain features or motor skills that are not supported by the NAO robot (e.g., jumping up and down or finger-counting), preventing a direct mapping from these recorded gestures onto the robot. For this reason, several gestures had to be reinterpreted, although the suggestions from the elicitation procedure were still used as a guideline. Figure 4.5 (left) displays an example of finger counting where such a reinterpretation had to take place: To depict the concept *four* using the robot’s fingers, we had the robot raise both hands showing two of its three fingers per hand by turning the wrist so that the thumb was hidden from view. Figure 4.5 (right) shows a gesture that could be translated more directly, without adjustments. The gestures for the robot were made using the Choregraphe tool that is provided with the NAO robot (Pot et al., 2009), which allows the designer to define key frames. The robot then interpolates between these key frames when producing a gesture.

An initial pilot evaluation with five verbs (out of the 34 target words) was conducted to validate whether the gestures’ comprehensibility, or how well the gestures matched the concepts they intended to describe, indeed influences how

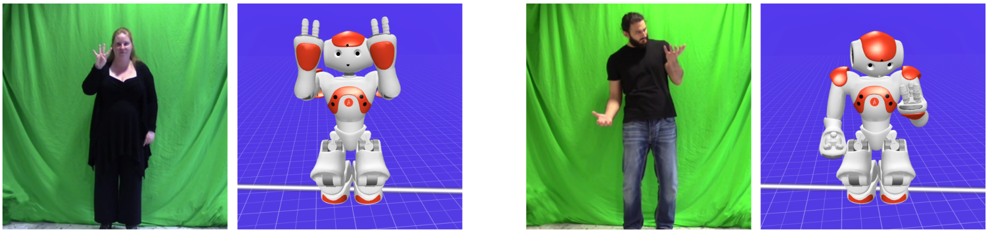


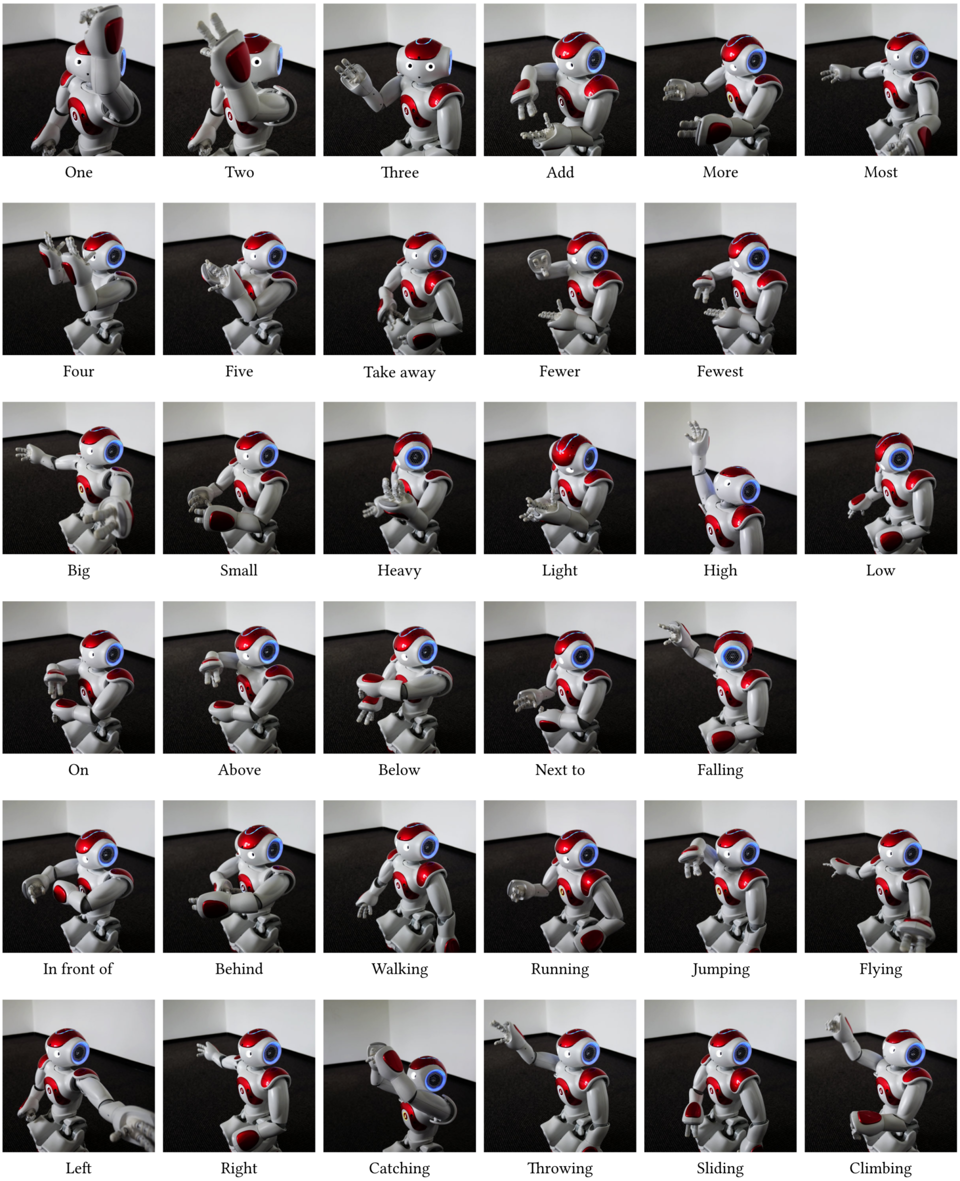
Figure 4.5: Examples of the translation of human-recorded gestures onto the robot for the concepts *four* (left) and *light* (right). Images used from the data of Kanero et al. (2018a) with permission.

well these gestures support tutoring by leading to improved learning outcomes. This was done by conducting a between-subjects study with children as participants ($N = 43$, $M_{age} = 5$ years, 9 months, $SD_{age}$ = 7 months), where the gestures were either performed by a robot or by a human tutor. The results, described in more detail in one of our project's deliverables[4], indicate that indeed the comprehensibility of an iconic gesture, as originally rated for human-performed versions by twenty adult participants, influences its effect on learning outcomes of children that use these gestures to learn English words, at least when this is measured by means of a receptive vocabulary task. No significant differences were found in a production task.

Before including them in the current experiment, the gestures were revised once more, especially taking into account the change in the robot's positioning relative to the child — in the original recordings, participants were standing and facing the camera, while in the experiment the robot was seated and placed at a 90-degree angle to the right of the child, changing the way gestures were perceived. Figure 4.6 shows photographs of all 34 gestures as they were used in the study, taken from the perspective of the learner.

There is a further distinction between the gestures that were designed for this study: Examples such as *running* use the whole body, where the robot actually "becomes" the runner (character viewpoint), while others such as *jumping* instead use one hand to depict an imaginary character or object that is jumping, also known as the observer viewpoint. Research has shown that younger children tend to use a larger gesture space, and perform gestures from the character viewpoint (as is the case with the *running* example) more often than smaller, imaginative gestures from the observer viewpoint such as the one for *jumping* (Sekine et al., 2018). This suggests that it could be better to use more gestures where the robot actually "becomes" the concept. However, this is not always possible given the robot's physical limitations.

[4] https://web.archive.org/web/20210415022714/http://www.l2tor.eu/effe/wp-content/uploads/2015/12/D7.4-Evaluation-report-storytelling-domain.pdf



One
Two
Three
Add
More
Most
Four
Five
Take away
Fewer
Fewest
Big
Small
Heavy
Light
High
Low
On
Above
Below
Next to
Falling
In front of
Behind
Walking
Running
Jumping
Flying
Left
Right
Catching
Throwing
Sliding
Climbing

Figure 4.6: Gestures for all of the 34 concepts in the study. Video recordings are available at: https://www.youtube.com/playlist?list=PLJreGGDWkgkqQUIsZXMgekMHP1T-_dfbU.

**Integration with the lesson content**

The built-in text-to-speech engine of the NAO robot is able to trigger events — such as performing a gesture — at specific points during the robot's speech output. This was used to align speech and gestures, as well as perform coordinated deictic gestures and shifts in the robot's gaze to guide the learner's attention. For the iconic gestures we introduced pauses in the robot's speech, such that the corresponding target word in the L2 would coincide with the stroke, the most salient part of the gesture. If possible, the pronunciation of the target word was timed for a moment with little to no movement, thereby minimizing any negative influences that motor noise could have on the audibility of the robot's speech. The robot then resumed talking in L1 after the gesture was completed.

## 4.4 Data collection

### 4.4.1 Procedure

**Group introduction**

Children were first introduced to the robot in a group setting. This was generally done with an entire classroom, including children that did not (yet) sign up to participate in the experiment, with the teacher also present. Previous research has shown that these group introductions reduce anxiety for subsequent individual interactions (Fridin, 2014; Vogt et al., 2017a). During the group session, the robot introduced itself as 'Robin' — a unisex name, leaving the robot's gender open to interpretation — and demonstrated some of its abilities, for example by performing several dances and by inviting the children to join in taking on a number of different poses. It also highlighted some of its limitations, for example by mentioning that it cannot hear very well, thereby instructing children to speak loudly. This was done so that researchers could clearly hear the children repeating after the robot during the lessons, allowing them to press the Wizard of Oz button on the control panel. Children were invited to shake the robot's hand, which helped them to bond with the robot.

**Pre-test**

The pre-test took place either on the same day as the group introduction, or shortly thereafter. Children were retrieved from their classroom one by one and brought to a separate, quiet room — often the same room in which they later interacted with the robot. They sat down at a table on which a laptop was placed, with a researcher sitting next to them. The researcher then walked through the different pre-test segments in a predefined order:

1. Peabody Picture Vocabulary Test (L1 vocabulary knowledge);
2. Translation task of the target words from L2 to L1;
3. Visual search task (selective attention);
4. Non-word repetition task (phonological memory);
5. Questionnaire measuring anthropomorphism.

Depending on the type of task, the child either answered verbally or pointed at items on the screen, while the researcher took notes on a paper sheet or pressed corresponding buttons on the keyboard. The researcher gave positively voiced neutral feedback to the child without indicating whether the answers given were correct or not. If the child did not know an answer to one of the tests, the researcher reassured them that this was not a problem and stimulated them to proceed with the tasks. After completing all segments the child was brought back to the classroom. The pre-test took approximately 45 minutes, and was recorded with a video camera.

### Lessons

Children who were assigned to one of the three experimental conditions took part in a total of seven lessons, which were scheduled so that children received two lessons per week, and never two lessons on the same day. As a result most children completed the lesson plan over the course of four weeks. The first lesson was planned at least one day after the pre-test.

The interactions were situated in a separate, quiet room at the school, where the robot was sitting on the floor next to the tablet. The child was collected from his or her classroom and invited to sit in front of the tablet, after which the researcher started the lesson using the control panel. While the child and robot completed the lesson together, the researcher was sitting behind the child to discourage the child from looking at him or her instead of the robot for feedback. If needed the lesson could be paused and resumed using the control panel. The end of a lesson was always marked by stars appearing and moving around on the tablet screen, after which the robot said goodbye and the child was brought back to the classroom. Each session with the robot took approximately 15–20 minutes to complete.

**Post-test**

The post-test was administered twice for each child, first an immediate post-test close to the last lesson (but at least one day later), and then a delayed post-test approximately 2–5 weeks after the immediate post-test. In both cases the child was retrieved from the classroom and brought to a quiet room. Similar to the pre-test, the child sat down at a table with the researcher sitting next to him or her. Using a laptop, the two translation tasks and comprehension task described in Section 4.4.2 were completed in the following order:

1. Translation from L2 to L1;
2. Translation from L1 to L2;
3. Comprehension task;
4. Questionnaire measuring anthropomorphism (only in the immediate post-test).

The researcher noted down the answers as they were given by the child. Each post-test took approximately 30–45 minutes to complete, and was recorded with a video camera.

### 4.4.2 Measures

Three different tasks were used to measure whether children learned and remembered the target words. This included two translation tasks, one from the L2 to the L1 and one from the L1 to the L2, to measure children's ability to freely produce translations of the target words. In both tasks the researcher would repeat a predefined sentence ("Wat is [word] in het [language]?" — "What does [word] mean in [language]?"), where the word was either in L1 or L2, and the language was either Dutch or English depending on the translation task. The pronunciation of the target words was made consistent by using recordings from a bilingual speaker of Dutch and English, which were embedded in a set of Powerpoint slides and then triggered by the researcher.

To measure children's comprehension of the target words in L2, we conducted a separate task where children were shown a set of Powerpoint slides, each slide containing three pictures or videos depicting a certain concept (Figure 4.7). A voice recording from a native speaker was played back every time a new slide was shown, asking "Waar zie je... [L2 word]" ("Where do you see... [L2 word]"), after which the child was asked to point at the corresponding picture or video. Depending on

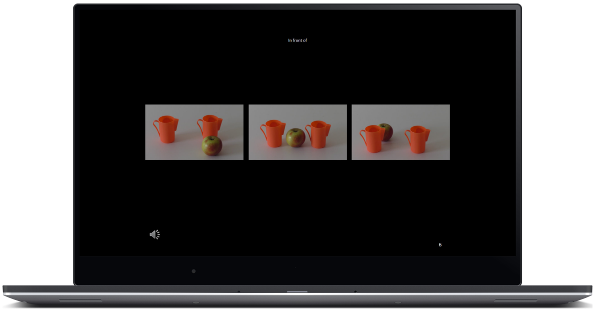


Figure 4.7: Example of the comprehension task (for the concept *in front of*) administered as part of the post-tests.

the target word, these stimuli would contain several physical objects, or a person performing a certain action. Because there is a relatively large probability that the children would guess correctly (33%), each concept was tested three times using different contexts, and shown with different distractor concepts (incorrect answers). However, because this would result in too many trials if all target words were included, we only tested 18 words, which were pseudo-randomly selected to include examples from all of the semantic categories (e.g., counting, measurement, movement verbs), and from all of the six lessons. Multiple versions were developed of both translation tasks and the comprehension task, in which the concepts were presented in a different order.

We further measured the children's receptive L1 vocabulary knowledge using the Peabody Picture Vocabulary Test (Schlichting, 2005), their phonological memory with a non-word repetition task (Chiat, 2015), and selective attention by means of a visual search task (Mulder et al., 2014). In addition, we investigated the extent to which children anthropomorphized the robot by means of a questionnaire (van den Berghe, de Haas, et al., 2021).

### 4.4.3 Analyses

In the current chapter, we conduct an in-depth analysis of the results, obtained using the measures discussed in Subsection 4.4.2, combined with an evaluation study of the comprehensibility of the robot's gestures. We focus on investigating how four different factors — comprehensibility of the gestures, age-based differences, differences between semantic categories, and gesture reenactment — may have affected children's learning outcomes. We will now present the analysis approach, followed by the results of these analyses in Section 4.5.

### Comprehensibility of the gestures

To investigate whether the meaning of the 34 final gestures included in the study was clear, we conducted an online evaluation study with 17 adult participants, 10 female and 7 male, with an average age of 21 years and 6 months ($SD = 2$ years, 8 months), recruited via convenience sampling. They were shown videos of all robot gestures, recorded from the same perspective as the photographs in Figure 4.6, in random order. Each video was between three to eight seconds long. Participants were asked to choose the concept belonging to the gesture they were just shown from a list of six possible answers. The incorrect answers were always the other concepts from the same lesson, to measure whether the 34 gestures were iconic enough to identify them within the context of the lesson in which they were used. The answers were also randomized for each trial. Lessons two (bakery) and four (fruit shop) contained only five target words in total, therefore the words *six* and *lifting* were added to these respective lessons as additional (incorrect) answers to ensure that the chance of guessing correctly was always the same.

Along with identifying the matching concept (binary scores, correct or incorrect), participants were asked to rate the clarity and naturalness of the gesture, both on a five-point scale ranging from 1. extremely unclear/unnatural to 5. extremely clear/natural. We then calculated the accuracy for each concept, which is the number of participants in the gesture evaluation study that correctly identified the concept divided by the total number of participants, resulting in a score from 0–1, as a measure of how comprehensible the matching gesture was. Correlation analysis (Kendall's tau-b, because of the relatively small sample size) was used to test whether the accuracy (as a measure of comprehensibility of the gesture), clarity, and naturalness are significantly correlated. In addition, we grouped the concepts into semantic categories, such as counting words and prepositions, based on existing language learning curricula[5]. Using paired samples t-tests, we tested whether there were significant differences between the semantic categories, in terms of the comprehensibility, clarity, and naturalness of the gestures.

To see whether the comprehensibility of a concept's gesture had an influence on children's learning outcomes for the English word belonging to that particular concept during the large-scale study, for each concept we calculated the score of the 54 children in the experimental condition where the robot used iconic gestures on

[5] See for example: https://www.gov.uk/government/collections/national-curriculum

the translation tasks. There were two translation tasks, from the L2 (English) to the L1 (Dutch) and from the L1 to the L2. Because there was a strong correlation between the two tasks, indicating that they both measure a similar language production skill, the scores on both tasks were averaged. This means that for each concept, the score of one child could be either 0 (incorrect on both tasks), 0.5 (correct on one of the two tasks), or 1 (correct on both tasks). For this analysis, we only included the experimental condition where the robot used iconic gestures, to focus on the relationship between gesture comprehensibility and the resulting learning outcomes when these gestures were used. Children's scores on the translation tasks were averaged across all children in the condition with iconic gestures ($N = 54$), to reach an average score for that particular concept (ranging from 0–1). We then compared the scores on both post-tests (immediate and delayed) for each concept to the rated comprehensibility of the gesture for that concept using correlation analysis. Note that the comprehension task of the post-tests only tested 18 out of the 34 target words, therefore we can only analyze the relationship between comprehensibility and post-test scores for these 18 words.

### Age-based differences between learners

To study the effect of the participating children's age on their learning outcomes, we ran the same analysis that was used in the original study (Vogt et al., 2019) to measure learning outcomes, but now with children's age at the time of the pre-test (in months) as a covariate. This analysis includes all four conditions so that we can investigate whether an observed effect of age applies to learning in general, or only when the robot uses deictic and/or iconic gestures.

The analysis is a doubly multivariate repeated measures ANOVA, with the translation scores (average of L2 to L1, and L1 to L2 translation tasks) and comprehension task scores as dependent variables, condition as independent variable, and age as covariate. The scores on the translation tasks were combined for all target words, which means that every participant had a score in the range of 0–34 (0.5 for each correctly translated word on one of the two translation tasks). The score on the comprehension task ranged from 0–54 (18 target words, 3 trials per word), where the chance of guessing correctly was 18 (33%), because every trial included the correct answer and two incorrect distractor items.

### Differences between semantic categories

For studying the differences between semantic categories, we included the tablet-only condition, where the robot was not physically present at all, and the two robot conditions (with and without iconic gestures), to see if the attention-guiding deictic gestures or the iconic gestures may have contributed to differences in children's learning outcomes for the different semantic categories. The 34 concepts for the translation tasks, and 18 concepts for the comprehension task, were divided into the same semantic categories used in the analysis of the gestures' comprehensibility (Subsection 4.4.3), and the post-test task scores were calculated for these semantic categories for the different experimental conditions. Scores on the translation tasks per child and per word were again either 0, 0.5, or 1, and for the comprehension task this was 0, 0.33, 0.66, or 1. These scores per child and per word were then averaged across children within the semantic categories, resulting in scores ranging from 0–1 for each category.

To check whether the differences between semantic categories were significant, we used a MANOVA with the scores on all six semantic categories, on the translation and comprehension tasks, as dependent variables (12 in total), and experimental condition (tablet-only, tablet + robot without iconic gestures, tablet + robot with iconic gestures) as independent variable. Furthermore, to test for an effect of age, in case differences occurred only for the older children in the sample, we ran the same MANOVA, including only the group of children who were at the average age of 5 years and 8 months or older (a *mean split*). This resulted in a subset of 38 children in the tablet-only condition, 32 in the tablet + robot without iconic gestures condition, and 29 in the tablet + robot with iconic gestures condition.

### Gesture reenactment

To investigate whether children that spontaneously reenacted the gestures benefited more from them than children who did not perform gestures themselves, we annotated these reenactment events and compared them with the children's learning outcomes. This was done by reviewing the recordings of the interactions of all children that were in the experimental condition where the robot used iconic gestures ($N = 54$), and noting down every occurrence of reenactment including the timestamp within the video and the concept that was reenacted. For feasibility reasons, this annotation was only done for the first lesson, with the underlying assumption that this would give a representative idea of how much reenactment actually took place during the entirety of the experiment. Furthermore, in the last two lessons the children were prompted to enact a number of action verbs, which in the experimental condition with iconic gestures essentially means that children were actively requested to reenact the gestures. Due to technical issues, the robot did not gesture during the first lesson for one of the children, therefore we had to exclude this child from the analysis, resulting in 53 observed sessions.

## 4.5 Evaluating the robot's gestures

In this section we present a detailed examination of the different factors that may have influenced the effectiveness of the robot's iconic gestures. We will first look at the comprehensibility of the gestures (Q1), followed by the role of age (Q2), differences between semantic categories of the target words (Q3), and finally the potential benefits of not only observing but also reenacting the gestures (Q4).

### 4.5.1 Comprehensibility of the gestures

Studies into human-performed gestures indicate that it is needed for iconic gestures to actually convey meaning, as meaningless or incongruent gestures do not appear to contribute to language learning and may in fact even have a detrimental effect (Kelly et al., 2009; Macedonia et al., 2011). We therefore investigated whether the meaning of the gestures included in the current study was clear by means of an online evaluation, and then compared these comprehensibility scores to the learning outcomes of children in the study to see whether the comprehensibility of the gesture of a particular concept contributed to learning the English word for that concept.

#### Evaluation study with adults

Appendix 4.A shows a full overview of the comprehensibility (accuracy) scores, and the ratings of clarity and naturalness, from the adult participants in the online evaluation study. Kendall's tau-b correlation was calculated to test the relationship between participants' accuracy in identifying the concept that was described by a gesture ($M = .72, SD = .09$) — the comprehensibility — and the rated clarity of the gestures ($M = 3.69, SD = 0.42$). This showed a significant medium correlation, $\tau_b = .37, p = .045$, where participants who rated the gestures as more clear also had a higher chance of matching this gesture with the correct answer. In addition, the correlation between gesture clarity and naturalness was significant, $\tau_b = .36, p = .043$, indicating that gestures that were rated as more clear were generally also rated as more natural. However, the correlation between comprehensibility and the rated naturalness of the gestures ($M = 3.48, SD = 0.35$) was not significant, $\tau_b = .35, p = .06$.

Table 4.2 presents a summary where we grouped the concepts by semantic categories. The lowest comprehensibility scores, .35 and .51, were found for counting words and comparatives, while operations and movement verbs had the highest comprehensibility scores: .97 and .88. Measurement words and prepositions received scores of .75 and .79. In the measurement words, the word *heavy* scored low (.29) compared to the other words in that semantic category, while *light* — which has a similar gesture — was generally identified correctly (.88). For the prepositions, the gesture for *on* scored especially low on comprehensibility (.47), compared to the other gestures in the same category. Using paired samples t-tests, we tested whether there were significant differences between the semantic categories. The results, which are presented in full in Appendix 4.B, show that there was a significant difference in comprehensibility between all semantic categories, except for measurement words and prepositions ($p = .30$). The clarity and naturalness ratings showed similar patterns to each other: They both differed significantly between counting words

Table 4.2: Comprehensibility (0–1), clarity (1–5), and naturalness (1–5) ratings for the gestures per semantic category (SD in parentheses). Chance level for comprehensibility is .17.

| Semantic category | Comprehensibility | Clarity | Naturalness |
|---|---|---|---|
| Counting<br>*One, two, three, four, five* | .35 (.11) | 3.00 (1.30) | 2.94 (1.13) |
| Comparatives<br>*More, most, fewer, fewest* | .51 (.23) | 3.22 (1.01) | 3.19 (0.92) |
| Operations<br>*Add, take away* | .97 (.04) | 3.03 (1.29) | 2.94 (1.18) |
| Measurement<br>*Big, small, heavy, light, high, low* | .75 (.24) | 3.92 (0.89) | 3.68 (0.87) |
| Prepositions<br>*On, above, below, next to, in front of, behind, left, right* | .79 (.15) | 3.89 (1.10) | 3.56 (0.98) |
| Movement verbs<br>*Falling, walking, running, jumping, flying, catching, throwing, sliding, climbing* | .88 (.11) | 4.11 (1.12) | 3.83 (1.10) |

and measurement words, prepositions, and movement verbs (all $p$-values < .001), between comparatives and measurement words, prepositions, and movement verbs (all $p$-values $\leqslant$ .007), between operations and measurement words, prepositions, and movement verbs (all $p$-values $\leqslant$ .01), and between prepositions and movement verbs (both $p$-values = .02).

In summary, the evaluation study of the gestures with adults shows differences in the comprehensibility (accuracy at the identifying the matching concepts), clarity, and naturalness, both between and within the different semantic categories. Particularly counting words and comparatives were often not correctly identified, while operations and movement verbs were relatively easy to recognize. Comprehensibility correlates with the rated clarity, but not the naturalness, of the gestures. Naturalness does correlate with clarity.

### Comprehensibility and learning outcomes

Figure 4.8 shows the comprehensibility scores — collected during the rating study with adults and discussed in the previous subsection — on the horizontal axis, and the children’s average scores on the translation tasks in the study on the vertical axis, for both the immediate (left) and delayed (right) post-tests. From these graphs we can identify three clusters, which appear for both post-tests:

1. High scores on the translation tasks, but low comprehensibility ratings — this cluster consists mainly of counting words, such as *four*;

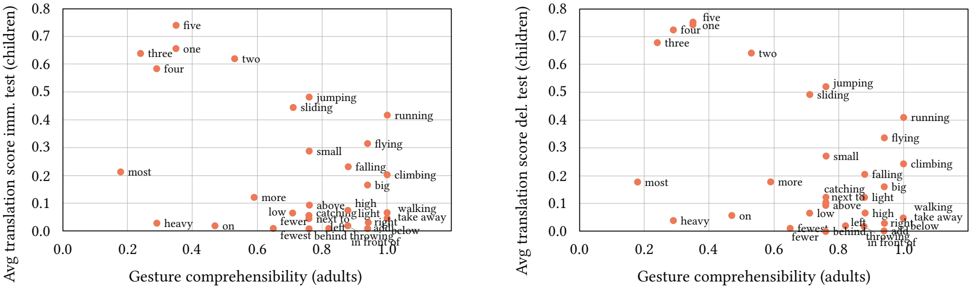


Figure 4.8: The individual gestures’ comprehensibility, in terms of mean accuracy by adult raters (horizontal axis), compared to the average translation scores of children in the condition with iconic gestures ($N = 54$) for these concepts (vertical axis). Scores per child were either 0 (no correct), 0.5 (correct on 1 translation task, L1->L2 or L2->L1), or 1 (correct on both tasks). Left: immediate post-test; right: delayed post-test.

2. Medium to high scores on the translation tasks, and high comprehensibility ratings — this cluster mainly includes movement verbs, such as *jumping*;

3. Low scores on the translation tasks, and medium to high comprehensibility ratings — this cluster includes most of the comparatives (e.g., *most*), operations (e.g., *take away*), measurement words (e.g., *heavy*), and prepositions (e.g., *behind*).

An analysis using Kendall's tau-b correlation shows that the correlation between the comprehensibility of the gestures, as rated by adults, and the scores of children participating in the condition with iconic gestures on the translation tasks was not significant for the immediate post-test ($\tau_b = -.19, p = .13$) nor for the delayed post-test ($\tau_b = -.20, p = .11$).

Figure 4.9 shows the same gesture comprehensibility scores on the horizontal axis, but now with children's average scores on the comprehension task on the vertical axis, for the immediate (left) and delayed (right) post-tests. Note that only 18 out of the 34 target words were included in this task. The results show a similar pattern for the comprehension task to the scores on the translation tasks, where children scored well on counting words and motion verbs. Additionally, children seemed to perform slightly better on some of the measurement words (*small*, *heavy*) and comparatives (*most*) on this task. Note that chance level for this score was 0.33. The Kendall's tau-b correlation between the comprehensibility ratings of the

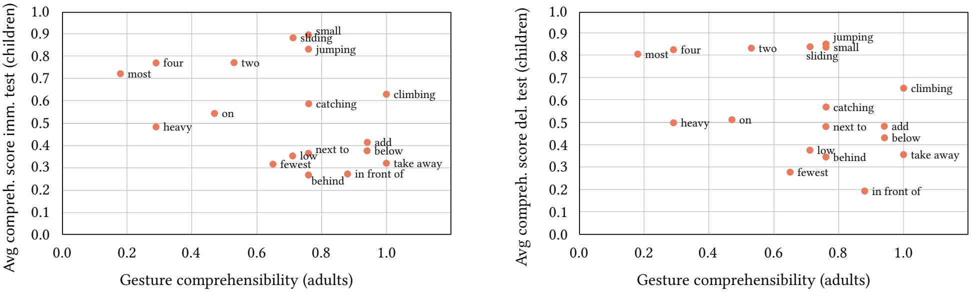


Figure 4.9: The individual gestures' comprehensibility, in terms of mean accuracy by adult raters (horizontal axis), compared to the comprehension task scores of children in the condition with iconic gestures ($N = 54$) for these concepts (vertical axis). Scores per child were either 0 (no correct), 0.33 (1 round correct), 0.66 (2 rounds correct), or 1 (all 3 rounds correct). Chance level is 0.33. Left: immediate post-test; right: delayed post-test.

gestures, and children's performance on the comprehension task was not significant for the immediate post-test ($\tau_b = -.17, p = .36$), nor for the delayed post-test ($\tau_b = -.17, p = .36$).

To summarize, with this analysis we do not find conclusive evidence that there is a relationship between the comprehensibility of the gestures, as measured with adults, and performance of children in the large-scale study on the post-test tasks. It appears that other factors, such as variation in difficulty of the concepts, play a larger role than the comprehensibility of the matching gesture.

### 4.5.2 Age-based differences between learners

Based on indications in existing research that the ability to perform and interpret (iconic) gestures develops during early childhood (Novack et al., 2015; Sekine et al., 2018; Stanfield et al., 2014), we explored whether age was a factor in children's learning outcomes during our study, with and without the robot's use of iconic gestures. Figure 4.10 shows the scores on the translation tasks of the immediate and delayed post-tests plotted against the participants' age in months at the start of the experiment.

A linear fit to these data shows a steeper curve for the experimental condition where the robot used iconic gestures, that starts at a lower score on the translation tasks for younger children compared to the other experimental conditions, while

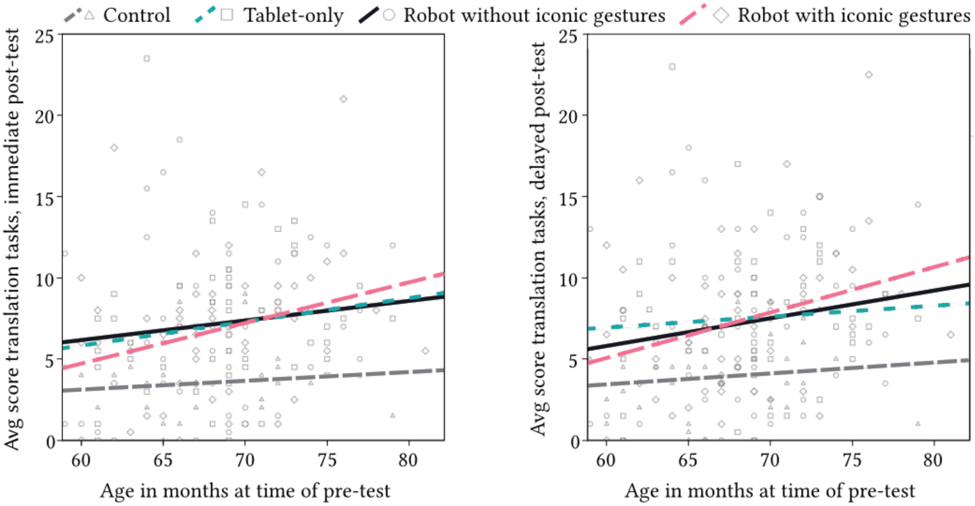


Figure 4.10: Linear fit to the post-test scores for the translation task per condition, by age.

it ends at a higher score than the other conditions for the older children in the study. This pattern does not emerge for the comprehension task, which is shown in Appendix 4.C.

A doubly multivariate repeated measures ANOVA, with translation scores (combined into one score for both translation tasks) and scores for the comprehension task as dependent variables, condition as independent variable, and children's age in months at the time of the pre-test as covariate, showed a significant effect of age for scores on the translation tasks, $F(1, 189) = 6.13, p = .01, \eta_p^2 = .03$, where older children in the study showed higher scores on the translation tasks of the post-tests than younger children. This effect was not significant for the comprehension task, $F(1, 189) = 1.24, p = .27, \eta_p^2 = .007$.

To further examine whether this effect holds for all experimental conditions, we split the dataset and ran the aforementioned ANOVA per condition, with the translation scores and comprehension scores as dependent variables, and age as covariate. This showed the same significant effect of age for scores on the translation tasks, but only for the experimental condition where the robot used iconic gestures, $F(1, 52) = 4.59, p = .04, \eta_p^2 = .08$. No significant effects were found for the comprehension task, nor for any of the tasks in the other three conditions (all $p$-values in range $[.26, .56]$).

The results of this analysis show that the older children in our study performed better on the translation (language production) tasks than the younger children, but only if the robot used iconic gestures while the children were learning the English words. Because this effect only shows in the experimental condition where the robot used iconic gestures, we postulate that older children may be better at understanding and making use of the robot's iconic gestures, compared to younger children. However, the effect of age should be interpreted with caution, because the effect size is relatively small.

### 4.5.3 Differences between semantic categories

Existing research suggests that iconic gestures for certain types of concepts (e.g., spatial concepts, motor events, or items that are relatively concrete) contribute more strongly to learning than gestures for concepts that are, for example, more abstract (de Nooijer et al., 2013; Hostetter, 2011; Wakefield et al., 2018). We therefore divided the English words into six semantic categories, and investigated whether there are any differences on average post-test scores between these categories, and if these can be attributed to the robot's use of gestures.

Table 4.3 shows the average scores of all children on the post-test tasks, on the immediate and delayed post-tests, for the three experimental conditions. The table shows no large differences between conditions for any of the semantic categories. To test whether there were any statistically significant differences, we conducted a MANOVA with the post-test scores on the six semantic categories, on the translation tasks and the comprehension task, as dependent variables (12 in total), and experimental condition as independent variable. This showed no significant effect of experimental condition on children's performance on the semantic categories for the immediate post-test (all $p$-values in range $[.11, .94]$), nor for the delayed post-test (all $p$-values in range $[.23, .99]$).

Because we observed an effect of age, where the older children appeared to benefit more from the iconic gestures than the younger children in the study, we also present the average post-test task scores on the semantic categories of children that were at the average age of 5 years and 8 months or older (a *mean split*). These results are displayed in Table 4.4. This table shows differences between conditions,

Table 4.3: Average translation and comprehension task scores (all 0–1) on semantic categories between conditions. T = tablet-only, NI = no iconic gestures, I = iconic gestures.

| | **Translation tasks** | | | **Comprehension task** | | |
|---|---|---|---|---|---|---|
| | T | NI | I | T | NI | I |
| **Immediate post-test** | | | | | | |
| Counting | .72 | .70 | .65 | .82 | .71 | .77 |
| Comparatives | .10 | .09 | .09 | .52 | .53 | .52 |
| Operations | .03 | .01 | .03 | .31 | .35 | .37 |
| Measurement | .10 | .09 | .12 | .56 | .57 | .58 |
| Prepositions | .03 | .04 | .03 | .42 | .41 | .36 |
| Movement verbs | .24 | .27 | .25 | .69 | .72 | .73 |
| **Delayed post-test** | | | | | | |
| Counting | .78 | .71 | .71 | .69 | .67 | .69 |
| Comparatives | .12 | .12 | .09 | .52 | .56 | .53 |
| Operations | .01 | .00 | .02 | .69 | .62 | .67 |
| Measurement | .09 | .10 | .11 | .64 | .65 | .63 |
| Prepositions | .03 | .04 | .04 | .52 | .55 | .55 |
| Movement verbs | .26 | .26 | .26 | .46 | .43 | .46 |

Table 4.4: Average translation and comprehension task scores (all 0–1) on semantic categories between conditions, for children that were at least the average participant age of 5 years and 8 months (*mean split*). Values in boldface are significantly higher than in the other experimental conditions. T = tablet-only, NI = no iconic gestures, I = iconic gestures.

| | **Translation tasks** | | | **Comprehension task** | | |
|---|---|---|---|---|---|---|
| | T | NI | I | T | NI | I |
| **Immediate post-test** | | | | | | |
| Counting | .77 | .75 | .74 | .84 | .74 | .86 |
| Comparatives | .10 | .09 | .10 | .54 | .59 | .53 |
| Operations | .02 | .02 | .04 | .28 | .33 | .39 |
| Measurement | .08 | .06 | **.16** | .55 | .55 | .61 |
| Prepositions | .04 | .04 | .03 | .39 | .39 | .35 |
| Movement verbs | .25 | .27 | .27 | .72 | .71 | .75 |
| **Delayed post-test** | | | | | | |
| Counting | .80 | .77 | .81 | .67 | .65 | .67 |
| Comparatives | .12 | .14 | .09 | .52 | .56 | .55 |
| Operations | .00 | .00 | .04 | .70 | .61 | .68 |
| Measurement | .08 | .06 | **.18** | .66 | .66 | .66 |
| Prepositions | .03 | .04 | .05 | .52 | .54 | .58 |
| Movement verbs | .27 | .26 | .29 | .46 | .43 | .49 |

particularly on the translation tasks for the measurement words, where children in the condition with iconic gestures scored higher than the children in both other conditions.

To test whether there were significant differences between conditions, the same MANOVA was conducted for this subset of older participants, which showed a significant effect of condition for the measurement words on the translation tasks on the immediate post-test, $F(2, 96) = 4.97, p = .009$, $\eta_p^2 = .09$, and for the translation tasks on the delayed post-test, $F(2, 96) = 5.85, p = .004$, $\eta_p^2 = .11$. For words related to operations, a significant effect of condition was found only for the translation tasks on the delayed post-test, $F(2, 96) = 3.60, p = .03$, $\eta_p^2 = .07$. No significant effects were found for categories other than measurement words on the translation tasks of the immediate post-test (all $p$-values in range $[.45, .96]$), and no significant effects were found for categories other than measurement words and operations on the translation tasks of the delayed post-test (all $p$-values in range $[.11, .85]$). Furthermore, no significant effects were found for any of the semantic categories on the comprehension task, neither for the immediate post-test (all $p$-values in range $[.11, .76]$) nor the delayed post-test (all $p$-values in range $[.23, .99]$).

For the measurement words, a post-hoc analysis using Bonferroni correction shows a significant difference on the immediate post-test between the experimental condition with iconic gestures and the tablet-only condition ($M_{dif} = 0.96, p = .047$), and between the conditions with and without iconic gestures ($M_{dif} = 1.21, p = .01$). There was no significant difference between the tablet-only condition and the condition without iconic gestures ($M_{dif} = 0.25, p = 1.0$). For the delayed post-test, a post-hoc analysis using Bonferroni correction shows a significant difference between the condition with iconic gestures and the tablet-only condition ($M_{dif} = 1.19, p = .017$), and between the conditions with and without iconic gestures ($M_{dif} = 1.39, p = .006$), but not between the tablet-only condition and the condition without iconic gestures ($M_{dif} = 0.20, p = 1.0$).

The post-hoc tests for the operations words on the delayed post-test showed no significant differences between the condition without iconic gestures and tablet-only condition ($M_{dif} = 0, p = 1.0$), between the condition with iconic gestures and the tablet-only condition ($M_{dif} = 0.17, p = .055$), or between the condition with iconic gestures and the condition without iconic gestures ($M_{dif} = 0.17, p = .07$). This is likely due to a floor effect, as shown by the .00 scores in the tablet-only condition and the condition without iconic gestures. Scores that are significantly different from the other experimental conditions have been marked in boldface in Table 4.4.

In summary, by comparing between experimental conditions we investigated whether the robot's physical presence, and its use of iconic gestures in particular, improved learning outcomes for specific semantic categories of words. When including all participants in the study, no differences between conditions were found for the semantic categories. However, after only including the older children in the study — those that appeared to be able to take advantage of the robot's gestures, as seen in Subsection 4.5.2 — we observe that the robot's iconic gestures were mostly beneficial to learning the measurement words (e.g., *big*), and they may have contributed to learning words pertaining to operations (*add, take away*) as well.

### 4.5.4 Gesture reenactment

In several studies that report a positive contribution of iconic gestures to learning, participants were asked to not only observe, but to also perform the gestures themselves (Cook et al., 2008; de Nooijer et al., 2013; Repetto et al., 2017; Tellier, 2005, 2008). We assume that this could lead to a stronger grounding effect of the new vocabulary in existing sensorimotor experiences.

In total, 37 out of the 53 children (70%) reenacted at least once during the first lesson. When children reenacted, they did this 13 times on average ($SD = 13$), out of a minimum of 60 gestures performed by the robot, depending on the number of times the robot had to repeat a task. Figure 4.11 shows the frequency distribution of how often children reenacted the gestures and the frequency distribution of how many different concepts (out of 6) children reenacted during the first lesson. To see whether the act of imitating the iconic gestures from the robot relates to learning outcomes, we calculated the Pearson correlation between number of reenactments in lesson one and test scores on the comprehension and translation tasks. Table 4.5 shows the results of this correlation analysis. The correlation was not significant for the translation tasks nor for the comprehension task, on both the immediate and delayed post-test. In Appendix 4.D we include a figure with each child's test scores on the vertical axis, and the number of times they reenacted during lesson one on the horizontal axis, showing no discernible pattern indicating a relationship between the number of reenactments during the first lesson, and children's learning outcomes. There was also no significant correlation between the children's age at the time of the pre-test, and the number of reenactments during the first lesson, $r = -.18, p = .19$.

Our investigation of spontaneous gesture reenactment shows that a relatively

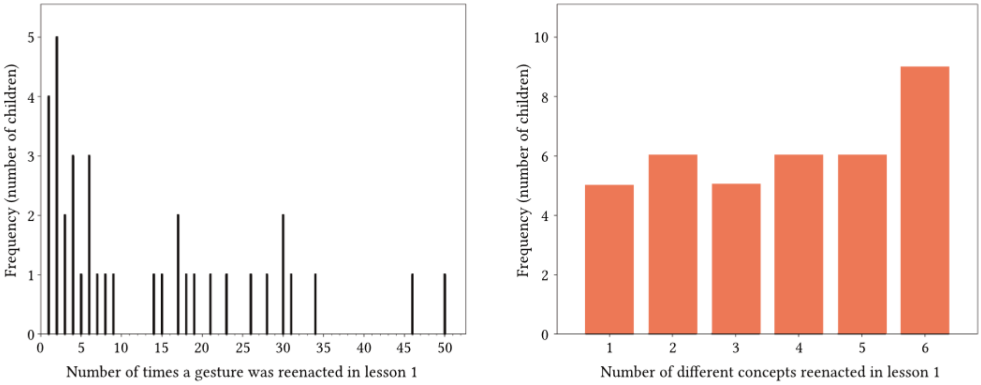


Figure 4.11: Left: Number of children (y-axis) that reenacted a certain number of times (x-axis) during the first lesson. Right: Number of children (y-axis) that reenacted a certain number of unique concepts (x-axis) during the first lesson. Only children that reenacted at least once are shown ($N = 37$; 16 did not reenact).

Table 4.5: Correlation between gesture reenactment and learning outcomes.

| | *r* | *p* |
|---|---|---|
| Immediate post-test, translation tasks | -.13 | .37 |
| Immediate post-test, comprehension task | -.09 | .53 |
| Delayed post-test, translation tasks | -.12 | .41 |
| Delayed post-test, comprehension task | -.17 | .24 |

large number of children reenacted the robot's gestures during the first lesson (70%), compared to our previous experiences with running similar studies. However, reenactment did not appear to affect learning outcomes, as there was no significant correlation between the number of reenactments in lesson one and the learning outcomes on the post-tests. In addition, the likelihood that a child in the study reenacted the robot's gestures did not appear to be linked to their age.

## 4.6 Discussion

Existing literature in gesture studies and human-robot interaction suggests that iconic gestures, performed by humans or by robots, are able to support second language tutoring. However, our previous study (Chapter 3) and the study that formed the basis of this chapter (Vogt et al., 2019) have shown mixed results, where in the case of our previous study the robot's iconic gestures did contribute to learning, while in the current study they did not. Therefore, in this chapter we set out to explore a number of factors that may influence the successful application of robot-performed gestures in second language tutoring. Concretely, we examined the importance of the design, and subsequent comprehensibility of the gestures (Q1), the age of the learners (Q2), differences between semantic categories of vocabulary words (Q3), and spontaneous gesture reenactment (Q4). In the following sections, we will address these subquestions, and infer guidelines for the design of robot-performed iconic gestures, focusing specifically on applications in (second language) education.

### 4.6.1 Design and comprehensibility of the robot's gestures

While reflecting upon the design of the robot's gestures, as well as their integration in the overall tutoring system, we have identified several differences compared to our previous study. First, the English vocabulary words included in the current study are more complex, diverse, and abstract than the animal names that were used previously. These words may have been more difficult for children to learn — as seen in the small number of new words learned in general — and the resulting gestures were less iconic than those from our previous study. A gesture for a concept such as *most* (shown in Figure 4.6), for example, will be more difficult to comprehend than a gesture that displays the act of riding a horse. In addition, the positioning of the robot may have affected the clarity of these gestures. While in the previous study the robot was standing across from the child, in the current set-up it was sitting close to the child, at an angled position. This limited the robot to only use its upper body, and it changed the perspective from which children were able to observe the gestures, which may have negatively affected their clarity. Concretely, we have seen that children misinterpreted gestures, as they were incorrectly mimicking them, for example by holding up their entire hand or showing three fingers for the word *two*. As a result of these factors, the gestures in the present research were likely more difficult to understand than those used in our previous study (Chapter 3), in which the gestures had a higher degree of iconicity, the robot was positioned facing the child, and the robot was able to use its full body to perform the gestures.

Although the gestures were designed based on recordings from an elicitation procedure, this procedure was conducted with adults rather than children from the same age group that would end up observing (and having to interpret) the gestures. Because children tend to perform gestures differently than adults do (Sekine et al., 2018), it is conceivable that they also understand gestures that were produced by their peers better than those produced by people from a different age group. In future work we propose to take a more iterative approach to the design of gestures, including more frequent evaluations and revisions — with the target demographic, in this case, children — before integrating the final versions into the tutoring interaction.

The online evaluation with adults of the gestures shows that there are differences in the comprehensibility of the gestures, both between and within the semantic categories. As we observed while conducting the robot experiment with children, the gestures for counting words were often misinterpreted because of the NAO robot's inability to move its fingers independently. We did not observe a clear link between the gestures' comprehensibility and children's performance on learning the corresponding L2 words. It would be an interesting avenue for future research to study more closely this link between the quality, in terms of comprehensibility, of robot-performed gestures and how this relates to learning outcomes. We would then consider conducting the gesture evaluation study with children belonging to the same age group that would end up interacting with the robot. However, it might be difficult for younger children to correctly identify abstract words such as *big* and *take away* without any context other than the gesture, and they cannot be asked to judge the clarity and naturalness of the gestures. We therefore intend to explore alternative ways of conducting these rating studies with children in the future, perhaps in a game-like setting and using the child's L1 to provide context.

From this evaluation study we also found a significant correlation between the comprehensibility and clarity of the gestures, as well as the clarity and naturalness, but not between naturalness and comprehensibility. Future research could look further into the nature of these relationships, to investigate how particular design aspects of gestures can be used to make the robot appear, for example, more human-like. It can be beneficial that a robot is perceived as human-like, as research has shown that this could, in turn, lead to the robot being perceived as warmer and more competent, which then leads to increased feelings of trust (Christoforakos et al., 2021). Robots that look and behave in a human-like way are generally also seen as more likeable, and are more easily accepted by the people interacting with them (Roesler et al., 2021). In addition, our previous research has shown that the degree to which a robot tutor is seen as human-like by children correlates with the children's learning outcomes (van den Berghe, de Haas, et al., 2021), which leads us to believe that a robot that is perceived as human-like could be more successful as a (peer) tutor than one that is perceived as a toy or an artificial entity. It would therefore be interesting to explore which aspects of the robot's gestures lead to higher ratings on naturalness and clarity.

Next to the design of the gestures themselves, and the limitations caused by the positioning of the robot, there are factors related to the integration of the iconic gestures into the intelligent tutoring system that could further explain why it may have been difficult for children to understand the gestures. For instance, the role that the tablet played within the overall interaction was smaller in our previous study compared to the current set-up. In Chapter 3, the robot was the instructor during a game of "I spy with my little eye" and children only had to select the correct image out of a number of answer options on the tablet. In the current experiment, children were asked to perform relatively complex tasks such as dragging objects in a three-dimensional virtual space. It is possible that children found these tasks to be more difficult, thereby drawing their attention away from the robot and its gestures. In addition, this could have increased cognitive load, resulting in less cognitive effort available to process the robot's gestures. An evaluation of the usability and user experience of the intelligent tutoring system also highlighted several issues that may have negatively affected the quality of the interaction, some of which occurred more frequently or even exclusively in the experimental condition with gestures (de Wit et al., 2019). Finally, the robot's gesturing interrupted the flow of the interaction. In order to time the motions so that the robot's pronunciation of the L2 words would coincide with the most salient part (the stroke) of the gesture, we introduced breaks in the robot's speech. Combined with the fact that 50–60 gestures were included in each lesson, this made the duration of the sessions substantially longer. As a result, the gestures had to maintain children's attention for a prolonged period of time. Research also indicates that a robot that gestures too frequently could be perceived as confusing and irritating (Pollmann et al., 2020), although this was found with adults and it is as of yet unclear how different gesturing frequencies by robots are perceived by children. Additionally, the robot performed the same gesture for a particular concept every time, so it is possible that children got bored with seeing an identical motion ten times. Although the same limitations apply to our previous study (Chapter 3), in the present study the interaction was more narrative-based, where the activities that the robot and child engaged in were linked to an overarching story line, compared to the more repetitive game of "I spy with my little eye" used in the previous study. During this previous study, the gestures were also repeated less frequently, and repetitions were spread out more over time.

### 4.6.2 Gestures and the effects of age

The fact that the older children in our study appeared to be able to understand and make use of the iconic gestures while the gestures seemed to have an adverse effect on younger children leads us to believe that either the gestures were too difficult or unclear for the younger participants in our study, or that younger children experienced some form of cognitive overload either due to the complexity of the interaction or the effort required to engage in learning second language vocabulary in combination with having to understand the gestures. Kennedy et al., 2015 also postulated that a robot's social behavior could lead to an increase of children's cognitive load, making it more difficult for them to focus on the task. Cognitive overload may have distracted the children from the (phonetic elements of the) robot's speech as it was practicing the L2 words with them.

It is worth noting that the effect size of age in the current study was relatively small, but so were the age differences (all children were approximately 5–6 years old).

To further investigate this effect, and to explore which factors may have affected the results, we have recently conducted a follow-up study where we returned to the original, single session experiment from our previous work (Chapter 3). We replaced the animal names with a more diverse set of concepts, and based the gestures on examples from a dataset of human-performed gestures — mostly performed by children and teenagers (Chapter 6). Interestingly, in this study we observed a similar effect where older children (six years old) did benefit from gestures, while younger children (four years old) appeared to experience an adverse effect, causing them on average to learn fewer words than children their age in the experimental condition where the robot did not use gestures. The effect sizes in this case were larger, which could be attributed to the broader age range of 4–6 years old, the design of the study (e.g., single session compared to longitudinal), or the different measurement instruments (comprehension task, measured as pre-test and post-test, compared to translation tasks and a comprehension task measured as post-test).

From both the current study and the follow-up study it appears that there is a certain (cognitive) development that occurs between the age of five and six, where children start being able to take advantage of the robot's gestures. Although literature indicates that we rely on gestures from a young age onward, it also shows that it takes time to fully understand and take advantage of them (Novack et al., 2015; Stanfield et al., 2014). Research by Stites and Özçalışkan (2017) further highlights that several aspects of gesture and speech change around the age of the participants in our study (5–6 years old). For example, they showed that children rely on gestures to support their speech when telling a narrative until the age of six, after which they start being able to use speech without support from gestures. It is therefore still possible that either the combination of foreign language learning and having to interpret gestures, or the multimodal interaction with a robot and a tablet may be too challenging for younger learners. This is further supported by a related study (van den Berghe et al., 2021b), where we found that children with better selective attention (as measured using a visual search task; Mulder et al., 2014) scored significantly higher on the post-tests if the robot used iconic gestures, compared to children with worse selective attention. It could also be the case that the children in our study differed in their ability to understand these two types of symbolic media — the robot's gestures and the depictions on the tablet screen (DeLoache, 2004). In future research we intend to run a gesture experiment with the robot but without a language learning component, in order to investigate whether this effect of age is indeed related to understanding the gestures, or to the cognitive effort needed to engage in the language learning interaction.

The effects of age on learning gain in the experimental condition with iconic gestures are only observed with results on the translation tasks, and not the comprehension task. This could either be because the gestures support children in acquiring a specific type of language skills (productive rather than receptive), or it could be due to the design of the tasks. Both our previous study (Chapter 3) and the recent follow-up study (Chapter 6) used only a (differently designed) comprehension task, and both found a positive effect of gestures on learning, either for all ages (Chater 3) or, similar to the translation task in the present study, with age as a covariate (Chapter 6). It is possible that the fact that only half of the concepts were included in the current comprehension task may have affected the quality of the measurements.

### 4.6.3 Differences between semantic categories

Existing literature indicates that gestures might be more effective at supporting learning of specific word types, such as spatial concepts or motor events (Hostetter, 2011), or verbs in general (Wakefield et al., 2018). We therefore compared the percentage of correct answers on the post-tests for words belonging to the different semantic categories between experimental conditions (presented in Table 4.3). However, it appeared that children in the experimental condition with iconic gestures did not learn different types of vocabulary words than children that were in the other experimental conditions (tablet only, or robot without iconic gestures). For the counting words, it is conceivable that children already knew these words before participating in the study, which would explain why they score well on these words even though the gestures were not recognized by adult participants in the comprehensibility rating study. This is further supported by the fact that there were no differences between experimental conditions, and by children's performance on the pre-test translation task (L2 to L1 only), where they generally scored well on the counting words. This does not apply, however, to the movement verbs, of which the gestures received high comprehensibility scores, and for which children had relatively high post-test, but not pre-test scores. Children in the experimental condition with iconic gestures did not score better than those in the other conditions, therefore these words in general seem to have been relatively easy for children to learn compared to other semantic categories. This may be supported by the fact that the children in all experimental conditions were asked to act out these movements during the lessons.

Because only the older children in the study benefited from the robot's use of iconic gestures, we performed the same analysis on the subset of 99 out of 194 children that were older than the average age of the entire group of participants (5 years and 8 months) — see Table 4.4. For this group we do see a difference in performance on the translation tasks: Children who interacted with the robot that performed iconic gestures knew more words from the measurement category (*big, small, heavy, light, high, low*), compared to children in the other experimental conditions. With the exception of *heavy*, these also had high comprehensibility scores, and above average ratings on clarity and naturalness in the gesture rating study (Appendix 4.A). Because the group of older participants within the experimental condition with iconic gestures is relatively small ($N = 24$), further research is needed to verify whether indeed gestures are more useful for certain types of concepts than others.

### 4.6.4 Gesture reenactment

Research on the potential benefits of not only observing but also reenacting or mimicking gestures is scarce, but initial findings indicate that this can indeed lead to better learning outcomes compared to merely observing others produce the gestures (Cook et al., 2008; de Nooijer et al., 2013; Tellier, 2005, 2008). To our surprise, in the current study 70% of the participants reenacted at least one gesture during the first lesson, without being prompted to do so. This is in stark contrast to our other studies with robot-performed gestures, where virtually no reenactment took place. Children might be more likely to imitate the robot's movements when it is positioned in a similar way to them: in this case both the child and the robot were sitting on the floor. The robot was also in relatively close physical proximity to the children. Because the robot was sitting, the gestures were generally smaller and limited to hand motions only, which may have made it easier and more inviting for children to reproduce the movements compared to more exaggerated, full-body movements. Furthermore, there was a short pose imitation game included in the group introduction of the robot, which the children may have remembered during subsequent interactions. Another potential reason for the more frequent reenactment of gestures is the word repetition task, where the robot requested the child to verbally repeat one of the English terms. This task was not included in our previous studies, and while introducing this task the robot would also gesture, which may have inspired the child to accompany his or her verbal repetition with a gesture as well. Generally speaking, it is possible that children in the current study formed a stronger relationship with the robot, compared to the previous study. This could be due to a multitude of factors (see, e.g., van Straten et al., 2020, for a review on child-robot relationship formation), such as the aforementioned physical proximity, and positioning the robot as a peer. Research suggests that familiarity with the demonstrator plays a role in whether children are likely to imitate behavior (e.g., Shimpi et al., 2013). Further investigation is needed to verify which aspects of the design of the interaction — e.g., the positioning of the robot, the inclusion of word repetition tasks, or relationship formation in general — can be used to stimulate gesture reenactment.

To investigate whether imitation of the robot's gestures had a similar positive effect on learning in our study as seen in literature with human-performed gestures, we annotated the number of reenacted gestures during the first lesson, all of which happened spontaneously as the robot did not ask the children to act out any concepts until the fifth lesson. We did not find a correlation between the number of times children mimicked a gesture from the robot during lesson one and their performance on the post-tests. However, since the effects of gestures on learning in general were small and only applied to the older children in the current study, we believe that more research into reenactment of robot-performed gestures is needed, with a direct comparison between observing and reenacting them, in order to come to clear conclusions. Children in all conditions (also those without iconic gestures) were invited to act out certain concepts in later lessons, particularly the movement verbs. They scored relatively high on these words on the post-tests, however with the current data we cannot be sure that this is caused by these enactments. These gestures were also found to be highly comprehensible, these words might have been easier to learn compared to other semantic categories, or children may have learned these words from other sources.

### 4.6.5 Strengths, limitations, and future work

With this work we continue our line of research into robot-performed gestures and their effects on children's acquisition of second language vocabularies. We focused on the specific domain of second language tutoring, and within this domain on a particular set of English words for which gestures were developed. Based on this study alone, we cannot conclude that our findings will generalize to a broader range of (educational) domains, user types (e.g., adults), and robot platforms without performing additional research. However, the results of the present study find support in existing research into human-robot interaction and gestures in general, which leads us to believe that this work is representative of the current state of social robots and robot-performed gestures. It is important to note that the original study, from which the data were used, was not designed with the aim to study the factors that are presented in this chapter. Rather, these explorations were conducted post-hoc. Therefore, in the future we aim to conduct several follow-up studies to investigate these individual factors in more detail.

The focus of this study was on children's learning outcomes, but there may be additional effects such as engagement with the robot or with the educational content, and perception of the robot that have not yet been analyzed. We believe that these aspects of human-robot interactions are important to consider, and are planning to include these in future work. Furthermore, we wonder to what extent the inclusion of a tablet device has affected our results, especially since the content that was shown on the tablet was designed specifically for this study. It would be interesting to conduct a similar language learning study either with existing educational software or without the tablet device present at all, and to measure how this affects the overall quality of the interaction, and the role that iconic gestures play in supporting this interaction.

The design of the robot's iconic gestures was based on examples recorded from human participants in an elicitation study. This is an improvement over designing these gestures using a researcher's frame of mind. However, these recordings did not take into account the physical limitations, nor the seated and angled positioning of the robot. In addition, the recordings and the evaluations of the robot's gestures were both conducted with adult, non-expert participants, even though children would end up interacting with the robot. It is possible that children have different preferences when it comes to gesture strategies, which were now not included in the design. Instead of iteratively refining the gestures based on multiple evaluations, due to time constraints we only evaluated the gestures once after the study had already taken place. While this still allowed us to control for the quality of individual gestures as a potential confound, it did not improve the quality of the gestures before they were used in practice. We have observed in related literature that a validation of the gesture's design prior to using them in a study often does not take place at all. Therefore, to further improve the quality of the gestures, we propose to conduct more frequent evaluations, and to include participants who have similar demographic characteristics to the intended target audience.

## 4.7 Conclusion

We report on the design of an intelligent tutoring system, which was used to investigate whether a social robot can be used as a second language tutor for young children, particularly focusing on the design of the robot's iconic gestures. In the original, preregistered analyses of the results of the study with this intelligent tutoring system, we observed no benefits of the robot's use of iconic gestures to learning. This is in contrast with our results from previous studies, and therefore in the current chapter we set out to investigate several factors that, based on literature, may play a role in the successful application of robot-performed iconic gestures to support learning. These factors included (1) the quality of the gesture's design (and subsequent comprehensibility of these gestures); (2) the age of the learner and how that may affect their ability to make use of the gestures; (3) differences in effectiveness of gestures depending on the concept that is being described; and (4) whether the learner reenacted or imitated the robot's gestures.

We found that, in the current study, gestures that were rated as more comprehensible by adults did not lead to better learning outcomes for children. The age of the participants did play a role in the experimental condition where the robot used iconic gestures: older children in this condition showed better learning outcomes compared younger children. The older children particularly benefited from gestures pertaining to measurement type words, such as *small.* Reenactment of the robot's gestures did not lead to increased learning outcomes in the current study.

This work contributes to the field of human-robot interaction by highlighting potential factors — gesture comprehensibility, age, types of concepts that are referred to, and gesture reenactment — that could play a role in the effectiveness of a robot's use of iconic gestures in an educational context. Based on our findings, we propose several improvements to the process of designing a social robot's iconic gestures, and integrating them in a (tutoring) interaction. In light of the present research and its promising outcomes, in future work we intend to conduct a study where we focus specifically on investigating these four factors in more detail.

* * *

*In Chapter 4, we provide an in-depth analysis of the results of our second study. Contrary to the first study (Chapter 3), we did not find a beneficial effect of the robot's use of iconic gestures on children's learning outcomes. We investigated whether four factors — children's age, comprehensibility of the robot's gestures, differences between types of words, and reenactment — influenced these results, and found that older children in our study benefited more from the robot's use of iconic gestures than younger children did. Based on this, we revisited Chapter 3 to investigate whether age was a factor in that study as well, but this turned out not to be the case. We therefore postulate that certain gestures, including those pertaining to animal names, might be easier for younger children to understand than gestures related to more abstract concepts.*

*The design of the robot's gestures in this study was based on an elicitation procedure with a small number of adult participants in the lab. To learn more about gesturing strategies and natural variation that might occur when people are invited to produce gestures for certain concepts, and to study differences in gesturing behavior between children and adults, in the next chapter we present a semi-structured elicitation procedure, using a game of charades with the robot, that was conducted in the field. This resulted in a set of recordings of naturalistic, human-performed gestures, that can be used to design gestures for the robot to perform in experimental studies, such as the one described in the current chapter.*

## 4.A Detailed Results of the Gesture Rating Study

Table 4.6: Comprehensibility, clarity and naturalness ratings for each gesture (SD in parentheses). Chance level for comprehensibility is 0.17.

| Concept | Semantic category | Compr. 0–1 | Clarity 1–5 | Natural 1–5 |
|---|---|---|---|---|
| One | Counting | 0.35 | 2.35 (1.00) | 2.41 (0.94) |
| Two | Counting | 0.53 | 3.41 (1.33) | 3.41 (1.00) |
| Three | Counting | 0.24 | 4.12 (0.86) | 3.88 (0.98) |
| Four | Counting | 0.29 | 3.06 (1.20) | 2.65 (0.86) |
| Five | Counting | 0.35 | 2.06 (1.03) | 2.35 (1.11) |
| More | Comparatives | 0.59 | 3.24 (0.90) | 3.12 (0.86) |
| Most | Comparatives | 0.18 | 3.12 (1.05) | 3.24 (0.90) |
| Fewer | Comparatives | 0.65 | 3.12 (0.99) | 3.18 (0.95) |
| Fewest | Comparatives | 0.65 | 3.41 (1.12) | 3.24 (1.03) |
| Add | Operations | 0.94 | 2.53 (1.37) | 2.53 (1.28) |
| Take away | Operations | 1.00 | 3.53 (1.01) | 3.35 (0.93) |
| Big | Measurement | 0.94 | 3.94 (0.83) | 3.65 (1.17) |
| Small | Measurement | 0.76 | 3.82 (1.13) | 3.71 (0.99) |
| Heavy | Measurement | 0.29 | 3.88 (0.93) | 3.65 (0.79) |
| Light | Measurement | 0.88 | 3.47 (0.72) | 3.41 (0.80) |
| High | Measurement | 0.88 | 4.24 (0.83) | 3.88 (0.78) |
| Low | Measurement | 0.71 | 4.18 (0.73) | 3.76 (0.66) |
| On | Prepositions | 0.47 | 3.71 (0.92) | 3.06 (0.90) |
| Above | Prepositions | 0.76 | 3.65 (0.93) | 3.18 (0.88) |
| Below | Prepositions | 0.94 | 4.35 (1.06) | 4.29 (0.69) |
| Next to | Prepositions | 0.76 | 3.12 (1.36) | 3.06 (1.03) |
| In front of | Prepositions | 0.88 | 4.06 (1.03) | 3.71 (0.92) |
| Behind | Prepositions | 0.76 | 3.35 (1.22) | 3.00 (1.00) |
| Left | Prepositions | 0.82 | 4.29 (0.77) | 4.06 (0.56) |
| Right | Prepositions | 0.94 | 4.59 (0.62) | 4.12 (0.78) |
| Falling | Movement verbs | 0.88 | 4.53 (0.72) | 4.29 (0.77) |
| Walking | Movement verbs | 1.00 | 4.88 (0.33) | 4.35 (0.93) |
| Running | Movement verbs | 1.00 | 4.47 (1.01) | 4.41 (0.62) |
| Jumping | Movement verbs | 0.76 | 3.12 (1.32) | 3.06 (1.09) |
| Flying | Movement verbs | 0.94 | 4.41 (0.80) | 3.65 (1.41) |
| Catching | Movement verbs | 0.76 | 3.41 (1.00) | 3.18 (1.01) |
| Throwing | Movement verbs | 0.88 | 4.65 (0.61) | 4.24 (0.97) |
| Sliding | Movement verbs | 0.71 | 2.71 (0.92) | 2.82 (0.81) |
| Climbing | Movement verbs | 1.00 | 4.82 (0.39) | 4.47 (0.51) |

## 4.B Differences between semantic categories

Table 4.7: Paired samples t-tests to test differences between semantic categories. * indicates significant difference.

| | $M_{dif}$ | $SD_{dif}$ | $t(16)$ | $p$ |
|---|---|---|---|---|
| **Comprehensibility (0–1)** | | | | |
| Counting–comparatives* | -0.16 | 0.31 | -2.14 | .048 |
| Counting–operations* | -0.62 | 0.24 | -10.59 | < .001 |
| Counting–measurement* | -0.39 | 0.26 | -6.23 | < .001 |
| Counting–prepositions* | -0.44 | 0.30 | -6.11 | < .001 |
| Counting–movement* | -0.53 | 0.27 | -8.14 | < .001 |
| Comparatives–operations* | -0.46 | 0.24 | -7.90 | < .001 |
| Comparatives–measurement* | -0.23 | 0.20 | -4.83 | < .001 |
| Comparatives–prepositions* | -0.28 | 0.24 | -4.72 | < .001 |
| Comparatives–movement* | -0.37 | 0.25 | -5.98 | < .001 |
| Operations–measurement* | 0.23 | 0.16 | 5.99 | < .001 |
| Operations–prepositions* | 0.18 | 0.23 | 3.23 | .005 |
| Operations–movement* | 0.09 | 0.15 | 2.50 | .02 |
| Measurement–prepositions | -0.05 | 0.19 | -1.05 | .30 |
| Measurement–movement* | -0.13 | 0.16 | -3.49 | .003 |
| Prepositions–movement* | -0.09 | 0.16 | -2.25 | .04 |
| **Clarity (1–5)** | | | | |
| Counting–comparatives | -0.22 | 0.66 | -1.38 | .19 |
| Counting–operations | -0.03 | 0.78 | -0.16 | .88 |
| Counting–measurement* | -0.92 | 0.56 | -6.83 | < .001 |
| Counting–prepositions* | -0.89 | 0.44 | -8.35 | < .001 |
| Counting–movement* | -1.11 | 0.44 | -10.39 | < .001 |
| Comparatives–operations | 0.19 | 0.89 | 0.89 | .39 |
| Comparatives–measurement* | -0.70 | 0.48 | -6.06 | < .001 |
| Comparatives–prepositions* | -0.67 | 0.57 | -4.82 | < .001 |
| Comparatives–movement* | -0.89 | 0.57 | -6.43 | < .001 |
| Operations–measurement* | -0.89 | 0.88 | -4.20 | .001 |
| Operations–prepositions* | -0.86 | 0.80 | -4.44 | < .001 |
| Operations–movement* | -1.08 | 0.71 | -6.30 | < .001 |
| Measurement–prepositions | 0.03 | 0.50 | 0.26 | .80 |
| Measurement–movement | -0.19 | 0.40 | -1.95 | .07 |
| Prepositions–movement* | -0.22 | 0.35 | -2.63 | .02 |

Table 4.7: Paired samples t-tests to test differences between semantic categories. * indicates significant difference.

| | $M_{dif}$ | $SD_{dif}$ | $t(16)$ | $p$ |
|---|---:|---:|---:|---:|
| **Naturalness (1–5)** | | | | |
| Counting–comparatives | -0.25 | 0.65 | -1.57 | .14 |
| Counting–operations | 0.00 | 0.74 | 0.00 | 1.00 |
| Counting–measurement* | -0.74 | 0.61 | -4.94 | < .001 |
| Counting–prepositions* | -0.62 | 0.44 | -5.73 | < .001 |
| Counting–movement* | -0.89 | 0.58 | -6.36 | < .001 |
| Comparatives–operations | 0.25 | 0.99 | 1.04 | .31 |
| Comparatives–measurement* | -0.49 | 0.52 | -3.86 | .001 |
| Comparatives–prepositions* | -0.37 | 0.49 | -3.10 | .007 |
| Comparatives–movement* | -0.64 | 0.68 | -3.89 | .001 |
| Operations–measurement* | -0.74 | 0.89 | -3.40 | .004 |
| Operations–prepositions* | -0.62 | 0.88 | -2.90 | .01 |
| Operations–movement* | -0.89 | 0.87 | -4.22 | .001 |
| Measurement–prepositions | 0.12 | 0.30 | 1.63 | .12 |
| Measurement–movement | -0.15 | 0.41 | -1.54 | .14 |
| Prepositions–movement* | -0.27 | 0.42 | -2.67 | .02 |

## 4.C Age and scores on the comprehension task

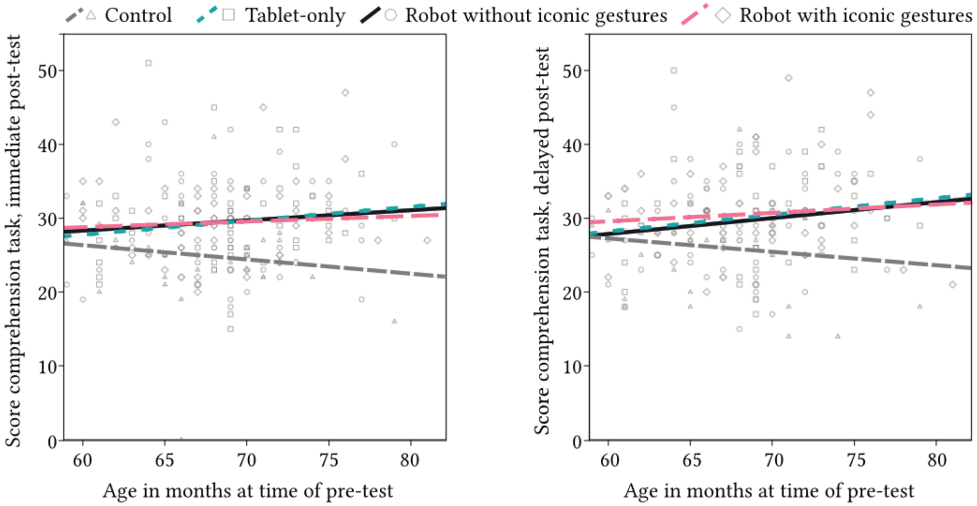


Figure 4.12: Linear fit to the post-test scores for the comprehension task per condition, by age (chance level is 18).

## 4.D Reenactment and learning gain

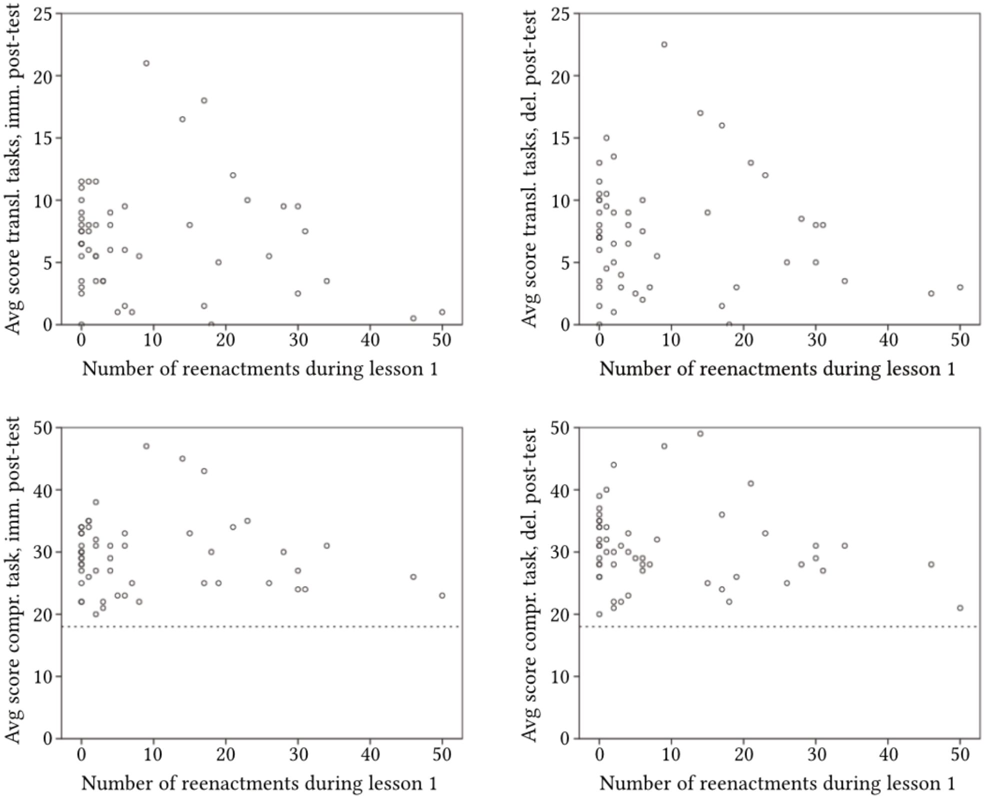


Figure 4.13: Post-test scores on the translation tasks (top) and comprehension task (bottom) plotted against the number of reenacted gestures in lesson 1. Chance level is 18 for the comprehension task.

## Chapter 5

# Studying Variation in Iconic Gesture: Introducing the NEMO-Lowlands Dataset

*This chapter is based on:* de Wit, J., Krahmer, E., & Vogt, P. (2020). Introducing the NEMO-Lowlands iconic gesture dataset, collected through a gameful human–robot interaction. In Behavior Research Methods, 1-18.[2] *Open practices:* The data and materials are available at https://osf.io/r59hj/ and at https://github.com/l2tor/NEMO-Lowlands-charades. The experiment was not preregistered.

## Abstract

This chapter describes a novel dataset of iconic gestures, together with a publicly available robot-based elicitation method to record these gestures, which consists of playing a game of charades with a humanoid robot. The game was deployed at a science museum (NEMO) and a large popular music festival (Lowlands) in the Netherlands. This resulted in recordings of 428 participants, both adults and children, performing 3,715 silent iconic gestures for 35 different objects in a naturalistic setting. Our dataset adds to existing collections of iconic gesture recordings in two important ways. First, participants were free to choose how they represented the broad concepts using gestures, and they were asked to perform a second attempt if the robot did not recognize their gesture the first time. This provides insight into potential repair strategies that might be used. Second, by making the interactive game available we enable other researchers to collect additional recordings, for different concepts, and in diverse cultures or contexts. This can be done in a consistent manner because a robot is used as a confederate in the elicitation procedure, which ensures that every data collection session plays out in the same way. The current dataset can be used for research into human gesturing behavior, and as input for the gesture recognition and production capabilities of robots and virtual agents.

[2] The Discussion section of this dissertation chapter has been updated with a footnote to introduce a study that was conducted using the dataset, after publication of the article.

## 5.1 Introduction

To support studies into non-verbal behavior, and in order to imbue robots and virtual agents with the ability to communicate with us in a human-like way, there is a need for structured, labeled, and large-scale datasets of human-performed gestures (Argall et al., 2009; Ortega & Özyürek, 2020). Ideally, these datasets contain gestures that are recorded in an ecologically valid way, and stored in a format that lends itself to automated analysis. Furthermore, it should be possible to collect additional data in a consistent manner, for example in order to include gestures for additional concepts or to replicate data collection in a new (demographic or cultural) context. With the aim to collect such a dataset of iconic gestures in a naturalistic setting, we developed a game of charades with a humanoid robot. This game was used to record a large number of iconic gestures from a diverse group of participants at the NEMO science museum and at the Lowlands Science event, as part of the Lowlands music festival. Both events took place in the Netherlands.

The resulting dataset of motion capture recordings for 35 different objects, such as animals and musical instruments, has a number of unique aspects that make it a valuable tool for studies and applications involving iconic gestures. First, it is a large-scale set both in terms of the number of unique recordings, as well as the number of participants that are included. Second, the participants were free to choose how they wanted to portray the concepts using silent gesture. Third, a broad range of demographic backgrounds — children and adults, several different cultures — is represented in the dataset. Fourth, to our knowledge no existing research has looked into the degree to which people tend to change their gesturing approach when an interlocutor fails to recognize their first attempt at depicting a concept. The current dataset provides support for first explorations into these repair strategies, and how often they were used. The combination of these four aspects has allowed us to capture different variations that are likely to occur in gesture production. This enables researchers to answer various research questions related to human-performed gestures, and factors that could potentially influence gesturing behavior.

The dataset contains two-dimensional and three-dimensional motion capture recordings of the participants performing the gestures. These are stored in a consistent format, which makes the set suitable for automated, large-scale gesture analysis, as well as various applications in the field of artificial intelligence such as gesture production and recognition by virtual agents and robots. Automatic gesture recognition is often done only for well-defined gestures, where the system knows what motion to expect. However, this means that people are limited in choosing their preferred way of depicting a concept using gestures. The current dataset allows researchers to explore whether it is possible to create recognition systems that can handle a variety of different representations for the same concept. An agent's gesture production capabilities can also be based on the recordings in our dataset, thus supporting studies into the added value of using data-driven gestures, and how comprehensible these are compared to manually designed gestures. Because the game of charades is made publicly available, it is possible to extend the dataset to include new concepts, or to record additional gestures in different cultures or contexts.

### 5.1.1 Gesture and interaction

Manual gestures (Kendon, 2004) are an integral part of our communicative abilities: they help guide the recipients' attention, and support the comprehension of information that is being conveyed in speech (Goldin-Meadow, 2005; Hostetter, 2011). They serve a purpose for the person producing the gestures as well, by helping them to be more fluent and rich in their speech (Cravotta et al., 2019; Hostetter, 2011). In this work we focus on iconic gestures, a specific subset that includes movements where the depicted shape is related to the concept that is referred to (McNeill, 1992). For example, an iconic gesture for the concept of a *bird* could consist of gracefully moving one's hands up and down repeatedly, as a reference to the act of flying. Iconic gestures in particular play an important role in supporting speech comprehension (Kelly et al., 1999), especially in noisy environments (Drijvers & Özyürek, 2017). Furthermore, people with certain impairments that prevent them from (fully) using or understanding speech, such as aphasia (language impairment due to brain injury), can benefit from gestures as a communicative and therapeutic device (van Nispen et al., 2018). Finally, research in the field of education has shown that iconic gestures can be used as a means of providing scaffolding to support the learning process (Alibali & Nathan, 2007). In light of this important role of iconic gestures in communication and education, with the current work we aim to provide a dataset and recording method to support further studies into the intricacies of gesturing behavior.

Because gestures are a natural and intuitive way for us to communicate with each other, researchers have started to explore whether we can use them to interact with machines as well (Karam & Schraefel, 2005). Recent technological developments enable everyday computer systems to track body posture and hand gestures in a minimally invasive fashion, which allows for the use of gestures as input device instead of using traditional controllers such as a mouse and keyboard (Lun & Zhao, 2015). This is especially relevant when these interactions involve artificial agents, either virtual or robotic, with whom we expect to be able to communicate by means of natural language (Bartneck & Forlizzi, 2004). Ideally, these agents should be able to understand the gestures produced by humans, as well as produce gestures of their own to support their social and communicative behaviors (Fong et al., 2003). We recently investigated whether gestures are able to support a robot's teaching efforts and found that children of 4–6 years old were more engaged with the interaction and showed higher learning gains when they interacted with a robot tutor that performed iconic gestures while teaching second language vocabulary, compared to one that did not use gestures (Chapter 3).

There are various methods — or modes of representation (Müller, 2014) — to describe a certain concept by means of iconic gestures. For example, one could gesture by outlining the physical shape of an object, such as the handle and bristle of a toothbrush, or by performing the act of using or interacting with the object: brushing our teeth. Although many concepts appear to have a default mode of representation (Dargue & Sweller, 2018; Masson-Carro et al., 2017; Ortega & Özyürek, 2016; Ortega & Özyürek, 2020; van Nispen et al., 2014; van Nispen et al., 2017), this is known to vary based on aspects such as the cultural background (Kita, 2009) or age of the performer (Jain et al., 2016; Masson-Carro et al., 2015; Sekine et al., 2018; Stites & Özçalışkan, 2017). The study by Sekine et al. (2018) showed that three-year-old children had a tendency toward using their entire body to represent the protagonist when retelling a story (character viewpoint), and they used a larger gesture space compared to adults. The adult participants instead performed gestures from the perspective of an outsider looking in (observer viewpoint), representing and manipulating the protagonist as a smaller, imaginative object. Even when performing the same gesture, Jain et al. (2016) observed that children of five to nine years old tend to produce faster and less coordinated motions than adults.

These variations in the way we depict concepts using gestures poses two challenges when attempting to imbue robots with the ability to understand and produce these motions. First, the robot-performed gestures are often designed by researchers using common animation techniques such as key framing. These researchers may not necessarily belong to the same demographic as the people that will end up interacting with the robot, and the robot's gestures might therefore fail to match the recipient's preferred modes of representation (Ortega & Özyürek, 2020), which could cause miscommunication. Second, a robot with social intelligence should also be able to recognize gestures performed by others, which are likely to include a number of variations for the same concept. Therefore, both the production and recognition of gestures by a robot would benefit from a data-driven approach, where many examples of people performing gestures are used to inform the robot's gesture production and recognition capabilities. There is a call for more data in the field of gesture studies as well (Ortega & Özyürek, 2020), in order to investigate whether patterns that we see on a smaller scale, e.g., regarding default modes of representation, can be generalized to a broader range of concepts or demographics. This ongoing research into human-performed gestures can be further supported by tools that have recently been developed to support automatic extraction of features such as size, velocity, and sub movements from three-dimensional gesture recordings (Pouw & Dixon, 2020; Trujillo et al., 2019), which enable analysis of gestures on a large scale. In order to improve the design of robot-performed gestures, and to support further studies into gesturing behavior, we have set out to collect such a dataset of three-dimensional recordings of human-performed gestures in a naturalistic setting.

These datasets can be collected in a number of different ways. For example, in recent work in the field of human-robot interaction, gestures were automatically extracted from natural interactions, such as recordings of TED talks (e.g., Ghosh et al., 2019; Hua et al., 2019; Shimazu et al., 2018; Yoon et al., 2019). These recordings were never intended to be used for this purpose, which means that the gestures that occur are naturalistic, but there is also no control over which (types of) gestures are performed. As a result, these gestures can be used for generating human-like co-speech gestures, but are less suitable for studying iconic gestures. The present work therefore focuses on the use of an elicitation procedure, which involves recording a number of participants as they perform gestures belonging to a predefined set of concepts. These concepts are presented to them one by one, either verbally or using visual cues. This method has also been used in the field of human-computer interaction, initially for the design of gesture interactions with a touch surface (Wobbrock et al., 2009), and subsequently for full-body gestures (e.g., Silpasuwanchai & Ren, 2014), also with children (Connell et al., 2013). The goal in the context of human-computer interaction is to reach consensus on the gesture that best describes a particular action within a computer system (Vatavu, 2019), such as shooting and reloading a gun in a videogame. Elicitation studies enable the collection of gesture datasets in a structured manner. It is possible to ask participants to perform examples of concrete motions (e.g., “claw like a bear”), but a more diverse set with different modes of representation can be collected by giving participants more general cues (e.g., “bear”). However, the data resulting from elicitation studies can be relatively unnaturalistic because participants are prompted to perform these gestures, often in a controlled setting, and they are aware of the goal and context of the study.

In order to obtain more naturalistic results, Eisenbeiss (2010) suggests the use of a semi-structured elicitation procedure, where the context is kept as natural as possible by having participants engage in a “game”, while still providing prompts to elicit certain responses. One example of a gameful approach is the director-matcher task. In this task a participant is assigned the role of director and is asked to describe a complex abstract shape to another participant, the matcher, who has to recreate this shape without having seen it (Krauss & Weinheimer, 1964). In gesture research, this method can be used to elicit a combination of speech and spontaneous gestures (e.g., Holler & Wilkin, 2011). This task can be considered an unstructured elicitation procedure, with little control over which exact gestures will be produced. Semi-structured and game-like approaches appear to be understudied in research. One example is Bartertown (van den Heuvel, 2015a), where participants engaged in a science-fiction game in which they were asked to communicate the appearance of certain primitive shapes to a virtual agent by means of gesturing. The recorded gestures were then mirrored by the virtual character and the participant was asked to confirm whether they were recorded correctly, and to re-do them if needed. Later in the game, other virtual characters performed gestures that were previously recorded from different participants and the current participant was asked to label these, essentially covering both the generation and labelling of data in one sitting.

To our knowledge, the potential use of repair strategies when there is a breakdown in non-verbal communication, both between two humans and between a human and a robot, has not yet been studied. However, we can find inspiration in the field of human-computer interaction, where mid-air (Walter et al., 2013) or touch gestures (Bragdon et al., 2010; Bragdon et al., 2009) can be used to trigger certain software commands. In this case, it takes time and multiple attempts for the user to explore which gestures are available, and to learn how they should be performed in order to trigger the correct functionality. Bragdon et al. (2009) found that a number of participants in their study either did not discover some of the available touch gestures at all, or they were unable to perform them in the proper way to trigger the functionality of the interface. This indicates a mismatch between the designer's expectations of the gestures that people will perform when interacting with their software, and the gestures that users actually come up with and the strategies they use to explore the space of potential gestures. We can apply the same principle to our studies in human-robot interaction: If we design the robot's gesture production and recognition capabilities solely on our own frame of reference, we are bound to introduce a certain degree of miscommunication. Therefore, it would be better to start by observing interactions, and then inferring common gesturing and repair strategies from these observations. Miscommunication can also occur when technology such as automatic speech recognition or, in our case, gesture recognition is not successful at recognizing the user's input correctly, a situation in which users can rely on multiple modalities for correcting these recognition errors (Suhm et al., 2001).

### 5.1.2 Existing gesture datasets

Several gesture datasets have been presented in literature, with various goals ranging from studies into human gesturing behavior, to applications related to artificial intelligence such as gesture recognition and gesture synthesis for virtual agents (e.g., Ortega & Özyürek, 2020; Sadeghipour et al., 2012; Vatavu, 2019). These sets differ in scale, in terms of the number of concepts included and the number of people recorded. Furthermore, different sensors were used to record the gestures, including traditional video cameras, depth sensors such as the Microsoft Kinect, and tracking devices that were held by or attached to the participants performing the gestures. These existing datasets can further be categorized by the elicitation procedure that was used, either (semi-)structured with specific cues, or unstructured where all of the gestures that were produced spontaneously during a broad task were recorded. An example of the latter approach is EGGNOG (I. Wang et al., 2017), where participants were given a collaborative task to recreate a structure out of wooden blocks from a picture. This resulted in a total of eight hours, collected over 360 trials with 40 participants, of naturally occurring gestures along with speech (for a subset of the trials). Another example is SaGA (Lücking et al., 2010), in which 25 pairs of participants were asked to perform tasks that involved giving directions and describing various scenes containing multiple objects. The resulting set contains recordings of speech and non-verbal behavior from 25 dialogues, including a total of almost 5,000 iconic and pointing gestures.

A literature review by Ruffieux et al. (2014) describes 15 datasets that were compiled specifically for developing and evaluating gesture recognition algorithms, which were collected using a structured elicitation procedure in a controlled setting. In most of the work discussed in this survey, the gestures do not refer to real-life objects, instead they are motions that were designed specifically to trigger certain actions during human-computer interactions (e.g., swiping to the right in the air to trigger the next song to play). Furthermore, participants were often given concrete prompts that already steered toward a particular aspect of the target concept, thus already implying a desired mode of representation, such as the aforementioned "claw like a bear" instead of just "bear". Only in the 3DIG dataset (Sadeghipour et al., 2012) participants were given the freedom to choose which representation technique (e.g., shape versus action) to use. This transforms the challenge of gesture recognition into being able to recognize any gesture that represents an object, rather than one specific motion. This form of gesture recognition is more realistic when communicating with (virtual) agents, where the focus lies on being able to understand which object is being described, regardless of individual differences in preferred gesturing strategy. In addition to their role in gesture recognition, such extensive and varied datasets can also be used for research into gesturing behavior in general. The 3DIG set contains recordings from a total of 29 participants, who were presented with ten primitive objects and ten complex objects such as *house* or *apple*. The aforementioned semi-structured elicitation procedure Bartertown (van den Heuvel, 2015a) also resulted in a publicly available dataset (van den Heuvel, 2015b), which includes three-dimensional gesture recordings of 36 participants each depicting 4 shapes, with 8 different shapes in total included. A recent example from the field of gesture research is the work by Ortega and Özyürek (2020), where 20 participants were asked to provide silent gestures for 272 different concepts across five semantic domains (manipulable and nonmanipulable objects, actions with and without objects, and animate entities), and were also given the freedom to choose their gesturing strategy.

Although the previously discussed datasets were all recorded with adult participants, there are datasets that include gestures performed by children as well. Vatavu (2019) published a set containing 1,312 whole-body gestures in total across 15 different concepts including objects such as flowers as well as actions such as climbing a ladder or turning around. These gestures were recorded from 30 children between the age of three and six. Children in this case were given concrete instructions on how to represent the concepts, for example to "Draw a flower in mid-air". The Kinder-Gator dataset (Aloba et al., 2018) contains recordings of 58 different gestures related to the categories warm-up, exercise, mime, and communication, such as "Motion someone to come here". These were recorded from ten children (aged five to nine) and ten adults.

Our survey of related work identifies several gaps in the datasets that are currently available. In most cases, participants were given concrete prompts for the types of gestures to perform, which makes these datasets unsuitable for studying individual differences in modes of representation. In addition, literature has found that gesturing strategies tend to differ between children and adults, however the only set from our survey of related work that includes both children and adults performing the same gestures is Kinder-Gator (Aloba et al., 2018). A limitation of this elicitation study is that the number of participants is relatively small (ten children and ten adults), they were given concrete prompts, and only few of the concepts elicited iconic gestures where the motion was semantically related to the concept being depicted. As a result, while this does support studies into quantifiable differences in motion characteristics (e.g., speed, size) between adults and children, it does not provide the variation needed to investigate differences in modes of representation. Finally, to our knowledge there is no iconic gesture dataset that includes the same participant performing a second gesture for the same concept, after they realize that the first example is not understood by the confederate. These second attempts would give insight into repair strategies that people tend to use when miscommunication occurs.

In our review of related datasets, we also found that generally none of the materials from the elicitation procedure that were used to collect the data are made available. This impedes potential future extensions of the datasets. In addition, the elicitation procedure relies on a human confederate, who has to follow a specific protocol. By having a robot perform this procedure instead, it is possible to replicate the data collection process in a consistent manner. In the present study, we aim to address the limitations of currently available iconic gesture datasets in two different ways: 1) by publishing a dataset that includes recordings from children and adults, who were free to choose their preferred mode of representation, and who were asked to perform a second gesture in case miscommunication occurred; 2) by making the game of charades publicly available, thereby allowing other researchers to further extend the dataset with different concepts, or in different cultures and contexts. Our dataset includes three-dimensional motion capture recordings from a depth camera, and two-dimensional motion capture data that were extracted from video recordings post hoc using an algorithm. Both formats have certain advantages and drawbacks, which will be discussed later in the chapter.

The next sections describe the game of charades with a robot that was used as elicitation procedure, followed by details regarding the technical implementation, and a description of the resulting dataset.

## 5.2 Gesture elicitation procedure

The game of charades was set up at the NEMO science museum in Amsterdam for two weeks in July and August, 2018, and at all three days of the Lowlands music festival, which took place on August 17–19, 2018. Visitors to the science museum and music festival were free to observe the study and, if they were at least five years old, could choose to volunteer as a participant. The study was carried out with approval from the research ethics committee of the Tilburg School of Humanities and Digital Sciences at Tilburg University. Participants, or their legal guardian in case they were younger than 16 years old, had to sign an informed consent form in order to participate, with which they also agreed that their data could be incorporated into the dataset. We also obtained verbal assent of all participants, and asked whether or not their interactions could be recorded on video in order to be able to extract two-dimensional motion capture data. These video recordings were optional, while the motion capture recordings from the depth sensor were required in order to participate.

### 5.2.1 Participants

A total number of 317 visitors to the science museum participated in the study, and 116 at the music festival. Due to children not finishing the game, or participants that took part in a demonstration of the system without wanting to have their data stored, we had to exclude five participants from the science museum. The total number of participants whose data were included, as well as their demographic information, is displayed in Table 5.1.

### 5.2.2 Materials

The experimental set-ups at the science museum and the music festival are shown in Figure 5.1. The system that was used in the experiment included a SoftBank Robotics

Table 5.1: Participant information.

| | NEMO | Lowlands | Total |
|---|---:|---:|---:|
| Participants | 312 | 116 | 428 |
| Gender | 157 male<br>149 female<br>6 unknown | 49 male<br>67 female | 206 male<br>216 female<br>6 unknown |
| Age (Y;M) | 12;11<br>*SD* = 10;7<br>11 unknown | 28;4<br>*SD* = 8;8<br>2 unknown | 17;2<br>*SD* = 12;2<br>13 unknown |
| Countries | 27 | 4<br>1 unknown | 28<br>1 unknown |

NAO V5 robot, a Kinect V2 for recording, a Microsoft Surface tablet as the interface for the participant and a control panel running on a separate laptop or computer for the experimenter. A Logitech C920 webcam was also included to capture video from which two-dimensional pose data were extracted after data collection was completed.

Thirty-five different concepts were included in the experiment for participants and the robot to depict. These were picked from the Bank of Standardized Stimuli (BOSS) containing photographs of a multitude of objects (Brodeur et al., 2014). Because we expected a substantial part of our participants to be younger children, we traced and colored the photographs to make them look more cartoon-like (Figure 5.2). The age of acquisition (Kuperman et al., 2012) was used as a guideline when choosing the concepts to ensure that the youngest participants (five years old) would be familiar with them. The concepts were divided into five different categories, with seven concepts in each category: animals, static objects, tools, musical instruments,

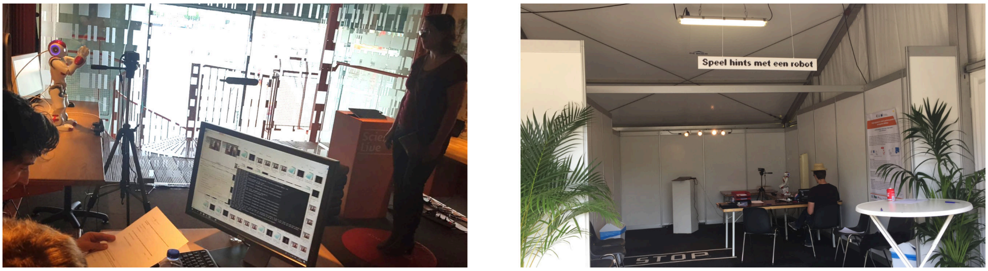


Figure 5.1: Photographs of the set-up at the NEMO science museum (left) and the Lowlands music festival (right).

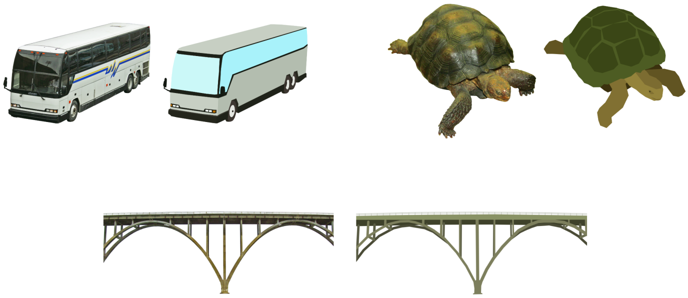


Figure 5.2: Three examples (*bus*, *tortoise*, and *bridge*) of photographs from the BOSS set and the corresponding traced images that were used in the game of charades — *tortoise* was also renamed as *turtle*.

and means of transportation. These categories were chosen in order to capture a diverse range of concepts, including both animate and inanimate objects, objects of varying sizes, and objects that afford different types of interactions (e.g., walking on a bridge, handling a toothbrush). To get a realistic idea of the robot's gesture recognition performance, several of the concepts were chosen to be similar to each other in terms of the default gesture we expected participants to use, such as car and bus, or xylophone and drum set. Table 5.4 at the end of this chapter contains an overview of all the included concepts.

### 5.2.3 Procedure

After visitors showed an interest in participating in the study, they were presented with a letter containing general information about the goals of the study, an explanation of the interaction with the robot (i.e., that they would play a game involving gestures), the nature of the recorded data (with a picture illustrating the output of the Kinect sensor), and details on the way their data would be collected and managed. To get an overview of what the game was like, visitors were also free to observe participants that were currently playing. After signing the informed consent form, their participant number was entered into the control panel. If the participant allowed their video to be recorded, a checkmark was set which enabled the system's video recording functionality. Additionally, participants could receive a link to a website with their own motion capture recordings. If they were interested in receiving this link, their e-mail address was entered into the control panel. The game was then started by the researcher by pressing a button on the control panel. The robot stood up and started "breathing" (shifting its weight from one leg to the other and swaying its arms slowly — a built-in feature of the NAO robot) to make it look more active and alive. It also blinked its eyes every five seconds by turning the LEDs off and on again. A language choice between Dutch and English was shown on the tablet, which affected the robot's speech as well as the labels for the items presented on the tablet.

The participant was invited to stand close to the tablet device so that they could operate it, and in front of the Kinect camera, which was moved approximately to the participant's shoulder height. The researcher then gave a short introduction to the game, indicating that the robot would only be able to see their upper body motion and instructing the participant to stand still with their hands pointing down at their sides when they were done gesturing. After choosing a language, the robot greeted the participant and explained the basics of the game to them. This was followed by a practice round, where the robot performed a prerecorded gesture to depict *glasses*, and the participant had to guess by selecting the corresponding image out of four different options (Figure 5.3).

Regardless of whether the participant guessed correctly or incorrectly, the game then proceeded to the second part of the practice round where the participant was asked to show a gesture for the object *ball.* After taking time to think of a way to depict the ball, the participant triggered a countdown by pressing the start button on the tablet, after which he or she could start performing the gesture (Figure 5.4). Participants were instructed to stand still after completing a gesture, which enabled the system to automatically detect when to stop recording. In a later version of the

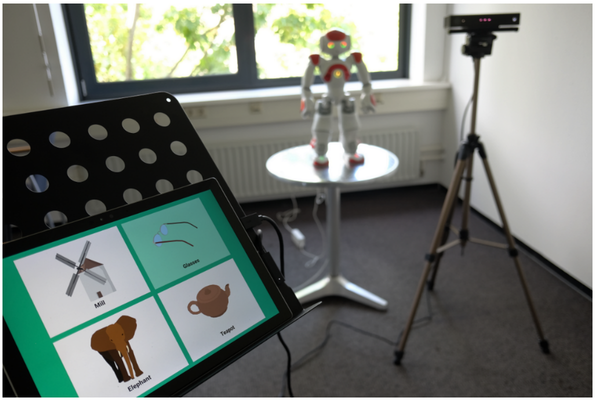


Figure 5.3: During the practice round, a participant guesses the gesture for *glasses* that the robot had just performed.

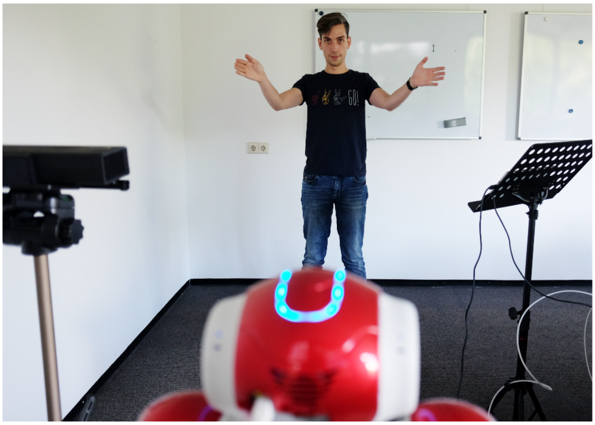


Figure 5.4: Second part of the practice round: The participant performs a gesture for *ball*.

system (used at Lowlands), there was also a button for the researchers to manually stop the recording. After the recording was stored, the robot tried to guess the gesture, which for this introductory stage was hard-coded to always be the correct guess regardless of the actual gesture that was performed by the participant.

After guessing the gesture, the robot displayed a top three of candidates for its guess along with a percentage showing how much confidence the robot had in that particular candidate. This step was included to give participants insight into the robot's thought process and reasoning behind its guesses. As with the other parts of the practice rounds, this was fixed and always showed the same three concepts with the same confidence values. All of the items used in the practice round were not part of the 35 concepts that make up the final dataset.

The participant then played five turns of the actual game, which were identical to the practice round except now with a selection of 10 out of the 35 included concepts — 5 to be depicted by the robot, and 5 by the participant. These concepts were chosen randomly, while ensuring that the number of total recordings across participants was equally distributed between the 35 concepts. The robot now based the gestures it performed on recordings from previous participants. In addition, it used a gesture recognition algorithm to try and identify the gestures performed by participants, and showed the actual top five candidates proposed by the algorithm. If the robot or participant guessed incorrectly a second attempt took place for the same concept. In many cases, this meant that the robot chose a different recording to perform for the concept. The four answer options on the tablet did not change, therefore participants had to guess from three items in the second round, because they already knew that one of the four original items was incorrect. The participants were also free to change their gesturing strategy for their second attempt (e.g., come up with a different mode of representation altogether, or repeat their previous gesture but then bigger or slower), although they were not actively asked to do so. The gesture recognition algorithm was purposefully implemented, even though it would mean an unequal number of repair attempts per concept, and per participant. We felt that it was important to offer a transparent and fair game experience to the participants, since we were working in two real-world environments. Furthermore, if participants would realize that the robot's guessing performance was controlled by us, they might not take the experiment seriously anymore, which would have negatively affected the quality of the recorded gestures. Figure 5.5 shows the information displayed on the tablet at various stages during the game of charades. The interaction with the robot lasted approximately ten minutes.

## 5.3 Technical implementation

In this section we present a general overview of the game of charades that was used as a semi-structured elicitation procedure. Additional details are available with the publicly available source code[2]. The implementation consists of several modules that communicate with each other using a local network connection. A key advantage of this architecture is that modules that have been developed in different programming

[2] https://github.com/l2tor/NEMO-Lowlands-charades

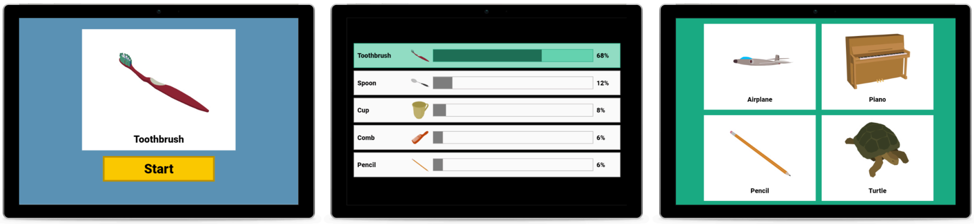


Figure 5.5: Screenshots of the tablet screen during the game of charades. Left: the participant's turn to perform a gesture for *toothbrush*; middle: when guessing, the robot shows its top five candidates, of which it will guess the first one (in this case *toothbrush*), the correct answer is highlighted; right: the robot just performed a gesture and the participant has to choose the matching item.

languages can still work together. The current system contains a combination of C# for Kinect, Javascript for the tablet interaction, and Python to drive the robot. In addition, each module is freely interchangeable as long as it sends the expected output to other modules and is able to handle the provided input. This means that different algorithms such as a better performing gesture recognition approach can easily be added in the future. In a similar vein, it is possible to support other robots or virtual agents as well as other recording devices without having to rebuild the entire system.

The current configuration uses a SoftBank Robotics NAO V5 robot, which is a commercially available and widely used humanoid robot. With 25 degrees of freedom it is more limited than humans in performing gestures. Most notably, it is unable to move its three fingers individually, so it is only able to open and close its hand in a gripping motion. In addition to the robot, the system requires a participant-facing tablet on which the game itself runs, and a computer where data can be stored and from which the researcher can control the experiment. We used a Microsoft Surface tablet for both the participant and the researcher. The human gestures were recorded using a Microsoft Kinect V2 depth camera, a device that was originally designed as an input device for the Xbox 360 gaming console but can be connected to a computer by means of an adapter. This device has since been discontinued but alternatives are available, including an updated version of the Kinect (Azure) which we aim to support with future updates to the source code. The robot and the devices for the participant and researcher were connected to a router via ethernet cables to ensure a stable connection. In the next two paragraphs we will briefly discuss the gesture recognition and production modules, two key components of the system.

### 5.3.1 Gesture recognition

To ensure that both the robot and the human participant were playing the game of charades fairly, both parties had to observe a gesture from the other player and then guess which concept it tried to describe. We therefore decided to implement an algorithm for the robot's gesture recognition capabilities. Because one of the potential use cases for our dataset was to train gesture recognition algorithms, this also enabled us to verify that the dataset was indeed suitable for this task. Finally, we could monitor the robot's gesture recognition performance as it interacted with people in a real-world setting and added new examples to the dataset.

Motion capture recordings such as the ones obtained from our game of charades are complex time series that describe three-dimensional locations of different joints (e.g., elbows, hands) over time. Therefore, even if two gesture recordings relate to the same concept and the performer used the same strategy to depict this concept, differences in speed, size of the movement, or the number of times a particular motion was repeated make it difficult to identify these similarities between gestures. A commonly used approach to compensate for these differences, particularly in speed, is dynamic time warping (e.g., Arici et al., 2014), which is able to match similar gestures even if they are not synchronized and move at different speeds. However, this method does not differentiate between motions that are crucial parts of the gesture, and the noise that stems from random movement or measurement errors during the recording of the gesture. It is also not robust to differences in participants' height, distance to the camera, or the size of the gesture, which may cause the joints' locations between two recordings to be far apart while the overall motion is in fact quite similar.

In order to distinguish between important movements and noise, and to also correct for differences in location due to the position or height of the participant, a preprocessing step is performed to identify salient features of the gestures, also known as *primitives* (Ramey et al., 2012). We based our approach on the work by Cabrera and Wachs (2017) by using the *inflection points* of the hands' motion trajectories, combined with peaks in the hands' position (Figure 5.6 shows a time series trajectory where inflection points and peaks are marked). Research suggests that inflection points are important features for humans to remember and reproduce gestures (Cabrera et al., 2017). To also take into account differences between participants' location and height and the size of the gesture, instead of the recorded absolute joint positions we use the positions relative to other joints. For example, we calculate whether the hand was in front of or behind, and above or below the shoulder. Cabrera and Wachs (2017) call the resulting sequence of inflection points and relative locations the *gist* of the gesture. One limitation that remains is that the same gesture could be performed at different positions relative to the body. A gesture for *ball* performed above the shoulders would therefore result in a different description than the same gesture performed in front of the body, below shoulder height.

The former preprocessing steps result in a feature vector describing salient points in the trajectory of the gesture. This feature vector consists of 14 dimensions that include the peaks near inflection points of the motion trajectory of the left hand relative to the left shoulder, the right hand relative to the right shoulder, the left

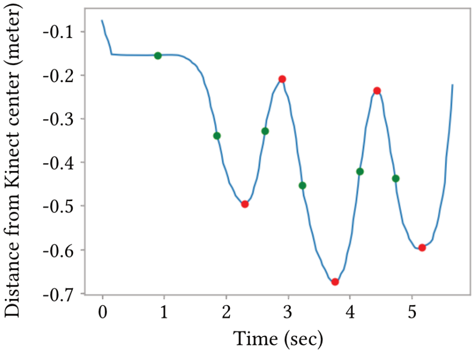


Figure 5.6: Inflection points (green) and peaks (red) of a motion trajectory.

hand relative to the right hand, and the spine at shoulder height relative to bottom of the spine (to measure bending/hunching). These peaks are all extracted from the X, Y, and Z trajectory, resulting in 12 dimensions. Each of these dimensions is a variable length text, which includes the location of the joint relative to the other joint (this is simplified by dividing the physical space into numbered quadrants), and whether at the inflection point the trajectory moved from convex to concave, from concave to convex, or whether it was a stationary point (+, -, or 0). Depending on the duration of the gesture and the number of salient points found within the trajectories of the limbs, one such dimension could contain between 0 and 39 salient points ($M = 1.95, SD = 2.47$ points). Each salient point is described by 2 characters of text: a quadrant identifier, and the type of inflection point. The last 2 dimensions of the feature vector are the percentage of the time the left and right hands were opened.

The next step is to find feature vectors of previously recorded gestures that are similar to that of the newly observed gesture. As a measure of similarity between gestures, we used the Needleman-Wunsch alignment score (Needleman & Wunsch, 1970), applied to the 12 dimensions of the feature vector separately. The similarity matrix is included with the published source code of the system. The difference in percentage of time that the hands were open was then subtracted from the similarity score. This helped the algorithm to distinguish between gestures that look similar if the hands are not taken into consideration, such as pretending to play the piano (open hands) and xylophone (closed hands). After calculating the alignment score between the new gesture and all existing ones in the set, the k-nearest neighbors algorithm (Altman, 1992) was used to determine to which concept the gesture was most likely to belong. This is done by taking the $k$ gestures with the highest alignment scores, in other words the $k$ recordings in the set that are most similar to the gesture we are trying to recognize. The value of $k$ was set to $\sqrt{N}/2$, where $N$ is the number of total recordings in the dataset. However, the maximum value of $k$ was set to 8 to ensure that the algorithm remained computationally feasible. This was determined empirically while developing the system, so it is possible that this is not yet the optimal value for $k$.

From the neighbors, the concept that occurred most often was chosen as the robot's guess (majority voting). For example, if the 8 closest matches included 4 recordings belonging to spoon, 3 to comb and 1 to toothbrush then the new gesture would be classified as spoon, and this is what the robot would then guess. If two concepts were tied (e.g., both 4 matches), the neighbor with the lowest similarity score was removed from the set of neighbors, and this process was repeated until there was one concept that had the largest number of matching neighbors.

All of the robot's guesses were logged while the system was deployed at the science museum and the music festival in order to get an overview of the gesture recognition performance and how this developed as more data were added. For both events, we initialized the dataset with three recordings for each of the 35 concepts, performed by one of the researchers. This was the starting point to which the system automatically started adding new recordings. Figure 5.7 shows the moving average, with an interval of 100 recognition attempts and exponential smoothing ($\alpha = .1$), of the robot's gesture recognition performance over time as it gained more data. Participants who did not want their data included in the analyses have been excluded. The average recognition rate was 17.7% at the NEMO science museum, and 21.0% at the Lowlands festival. Chance level is approximately 2.9% — 1/35 for first attempts, and 1/34 for second attempts at guessing. The Lowlands set contains less data because the system only ran for 3 days at that location, compared to 14 days at NEMO.

### 5.3.2 Gesture production

The recorded gestures do not contain any visual information, essentially turning the performer into a stick figure. This results in a loss of information compared to regular video recordings: context and facial expressions are missing, and subtle motions may not have been picked up by the Kinect camera. A further loss of information occurs when trying to automatically translate these recordings to a robot with

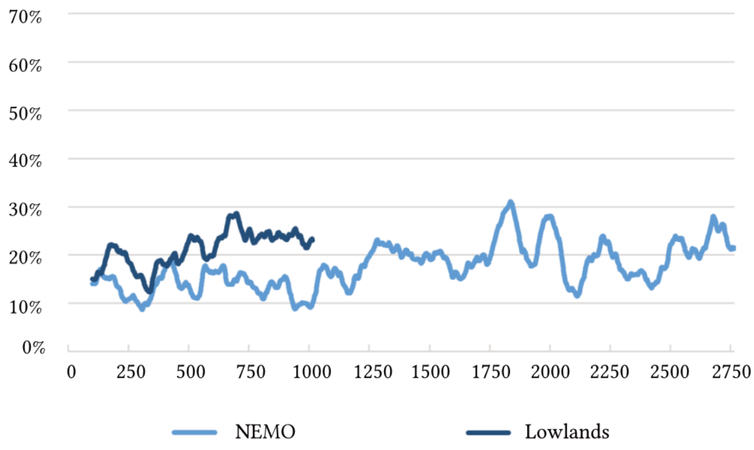


Figure 5.7: Moving average of the percentage of correct guesses by the robot.

fewer degrees of freedom, less smoothness in its motion, and a smaller reach than a human. However, if this automatic translation were to work while preserving the comprehensibility of the gestures, the robot would have the possibility to imitate human-performed gestures, so that the gestures no longer have to be designed by hand.

To measure the comprehensibility of the gesture recordings and the impact of the loss of information resulting from the recording and translation steps, we had the robot directly use gestures that were previously recorded from other participants. We used an existing implementation to translate the joint locations as they were recorded by Kinect into the yaw, pitch, and roll values needed by the robot (Suay & Chernova, 2011). Because it is not possible for the robot to perform certain motions as fast as a human can, the recordings were slowed down and then sampled at 300ms intervals. In addition, there were recordings where the system did not register that the gesture had ended, and thus also captured noise at the end. Therefore, only a maximum of ten seconds of the recordings were performed by the robot.

Each recording had a *weight* assigned to it, which started at 0 and was updated after the robot had performed this particular recording to a participant. If the participant guessed the corresponding concept correctly, the system increased the weight of this gesture. If the participant chose an incorrect answer, the system decreased the weight. These weights were then used when deciding which recording to use next. To make it easier for the participant to guess a gesture correctly, the robot could perform the recording with the highest assigned weight (the one that had been guessed correctly most often in the past). On the other hand, the robot could also avoid the highest scoring example and explore alternatives instead. To get diverse ratings while still providing participants with a good chance to win, in the current set-up we implemented a 60% chance that the "best" example would be used (exploitation), and a 40% chance that any other recording would be performed by the robot (exploration). Although it would have been possible to ensure that each gesture would receive an equal number of ratings, we opted for this exploration-exploitation approach to lower the difficulty for participants to win the game, and to automatically filter out incorrect or unclear gestures (noise).

Similar to the automatic gesture recognition performance, it is possible to see from the log files how well participants were able to recognize gestures performed by the robot by measuring how often participants guessed a gesture correctly. Figure 5.8 shows the moving average, with an interval of 100 recognition attempts and exponential smoothing ($\alpha = .1$), of participants' guessing performance. On average, participants guessed correctly 41.9% of the time at NEMO, and 50.3% of the time at Lowlands. Chance level in this case is between 25% (first attempt at guessing, four possible answers) and 33.3% (second attempt at guessing, three possible answers).

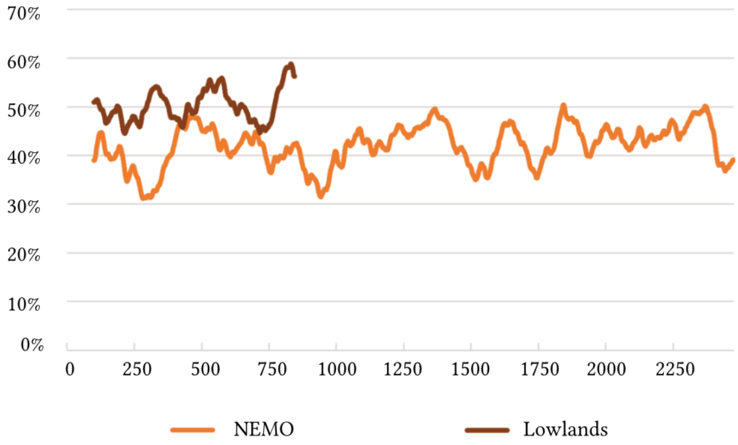


Figure 5.8: Moving average of the percentage of correct guesses by participants.

## 5.4 Description of the resulting dataset

After deploying the system at the NEMO science museum and the Lowlands music festival, the resulting data were cleaned and then published on the Open Science Foundation[3] as supplementary materials to this chapter. The dataset includes metadata describing the participants' age, gender, and country of residence, as well as the three-dimensional gesture recordings from the Microsoft Kinect V2 and the two-dimensional gesture recordings that were extracted from videos of participants that gave permission to have them recorded. These recordings are grouped in folders, one for each of the 35 concepts. Each filename contains the participant number, and whether this was a first or second attempt at performing the gesture. We have published the data for each of the two data collection locations separately, although they can easily be combined into a larger set by merging the folders with each other as the 35 concepts were the same between locations. The first character of the participant numbers can then still be used to tell entries from the different locations apart (N = NEMO, L = Lowlands).

Also included in the dataset are log files of all the sessions, which document the interactions that occurred (e.g., which exact gestures the robot performed, and all guessing attempts by the participants and the robot), as well as Python scripts that can be used to visualize (play back) the recordings.

### 5.4.1 Three-dimensional recordings

The Kinect V2 depth sensor is able to track the position of 25 different body joints (e.g., head, hips, hands, feet) at 30 frames per second. For each recording, we stored the estimated X, Y, and Z position of the 25 joints through time in a comma-separated (.csv) text file, with one line for each timestep. The Kinect uses the center of its sensor as the origin (0, 0, 0), and measures joint positions by their distance in meters from this origin. This means that the value of X increases as you move to the left of the sensor (from the perspective of the sensor, facing the participant), Y increases as you move up from the sensor, and Z increases as you move further away from the sensor. As a result, what is reported as the right shoulder was in fact the participant's left shoulder, as seen from the Kinect sensor. In other words, these recordings will be mirrored by default when played back. Figure 5.9 shows a frame from three different recordings for the concept *bridge*, visualized from the comma-separated file using one of the Python scripts included with the dataset. Note that not all participants

[3] https://osf.io/r59hj/

were standing far enough away from the sensor for it to be able to capture their entire body, hence the positions of their lower joints (e.g., knees and feet) could not be tracked.

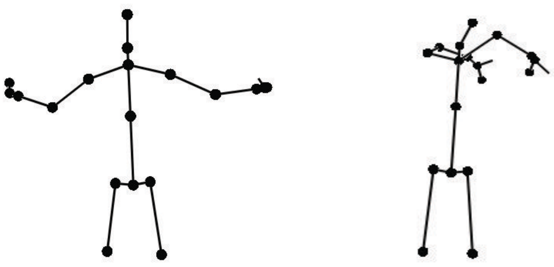


Figure 5.9: Three recordings of *bridge*, showing different ways of depicting this concept using gesture. From left to right, the focus is on the bridge surface, the arches, and opening of a drawbridge. The leftmost example is performed by an adult (24 years old), while the other two examples are by children (9–10 years old).

In addition to the 25 joint positions, the system stored joint orientations (in X, Y, Z, W), but these appear to be redundant with the joint positions and are therefore not used in our current implementation. The estimated face orientation — an indication of where participants were looking — was also added, which has been converted into pitch, yaw, and roll values. Finally, although the sensor cannot track individual fingers, it is able to determine whether the participants' hands are open (1), closed (0), or whether this is unknown (-1). This information was also added for each hand at every timestep, along with a confidence value indicating how sure the system was that the hand was in fact opened or closed (*Low* or *High*).

The total number of unique three-dimensional recordings is 3,715. Table 5.2 shows how many gestures are in each subset, and how many of the recordings were first or second attempts from the same participant. Table 5.4 at the end of this chapter provides a more detailed overview of the number of recordings per concept for each subset.

### 5.4.2 Two-dimensional recordings

Out of the 428 participants in our study, 367 gave permission to also have their gestures recorded on video. A Logitech C920 webcam was used, which captured the gestures at 25 frames per second. After data collection had completed, we first

Table 5.2: Number of three-dimensional gesture recordings per location, divided into first and second attempts.

| | First attempts | Second attempts | Total |
|---|---:|---:|---:|
| NEMO | 1,512 | 1,198 | 2,710 |
| Lowlands | 561 | 444 | 1,005 |
| Total | 2,073 | 1,642 | 3,715 |

corrected the video recordings for the camera's lens distortion, and then extracted motion capture data using OpenPose (Cao et al., 2017). This resulted in a similar data file to the three-dimensional Kinect recordings, including the positions of 25 body joints through time, but without depth information (the Z-coordinate). The X and Y coordinates in this case were measured in pixel locations within the video frame, which had a resolution of 1280x720 pixels, with the top left corner of the frame as the origin (0, 0). In these data the left shoulder refers to the participant's viewpoint, so it is actually positioned further to the right than the right shoulder (which shows up on the left side of the video recording). Contrary to the three-dimensional recordings, these will therefore not be mirrored when played back. In addition to the 25 body joints OpenPose is able to track 21 keypoints on each hand (i.e., finger joints), and 70 points describing the outline and features of the face. This approach is therefore able to extract several details from video that are missing from the three-dimensional recordings, such as finger movement or facial expressions. Figure 5.10 shows a comparison between recordings using Kinect, and the results of running OpenPose on video recordings of the same gesture. Similar to the three-dimensional recordings, lower parts of the body such as the feet were often obscured from view and could thus not be tracked.

Because not all participants (367 out of 428) gave permission to have their gestures recorded on video, only 3,269 out of the 3,715 gestures could be analyzed using OpenPose. Table 5.3 shows how these are distributed between the two locations, and how many first and second attempts from the same participant were included. The number of two-dimensional recordings for each concept is listed in Table 5.5 at the end of this chapter.

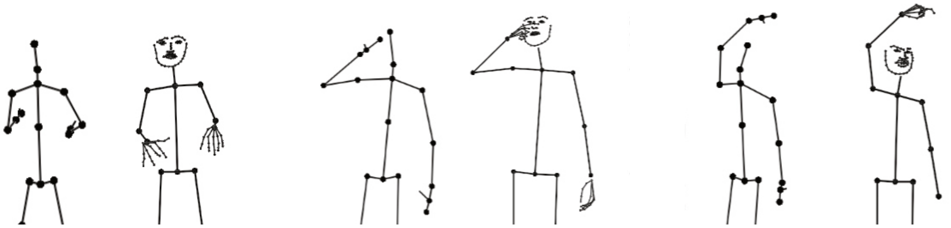


Figure 5.10: The two-dimensional and three-dimensional versions of three recordings, highlighting the advantages of having detailed hand and face motion. From left to right: *piano* with extended fingers, performed by a child (5 years old); *pig* by pushing the nose upward with the index finger, performed by an adult (26 years old); *stairs* with a walking motion by moving the index and middle fingers, performed by a child (10 years old).

### 5.4.3 Data cleaning

Because the recording of each gesture was started by the participant, and finished after the system detected little to no hand movement for a certain amount of time, each gesture was automatically isolated and stored in the folder belonging to the right concept, in its own file, with the filename including the participant number and whether it was a first or second attempt. We have reviewed all of the recorded gestures, and identified 34 recordings from NEMO, and 3 from Lowlands in which no movement resembling an iconic gesture took place. These were removed from the dataset.

Although the system tried to automatically isolate the gestures, there were cases where the system prematurely detected the end of a gesture and therefore the recording was cut short. There are also examples where the system did not manage to detect the end of the gesture due to too much idle movement by the participant. Because a certain degree of noise is to be expected once interactions such as these are deployed in a naturalistic setting, we have not edited the recordings to remove

Table 5.3: Number of two-dimensional gesture recordings per location, divided into first and second attempts.

| | First attempts | Second attempts | Total |
|---|---:|---:|---:|
| NEMO | 1,284 | 1,013 | 2,297 |
| Lowlands | 541 | 431 | 972 |
| Total | 1,825 | 1,444 | 3,269 |

these extraneous movements. The recordings might also contain participants looking at, or trying to interact with the tablet device as they double-checked the concept they were asked to perform, or if they did not realize that the recording had already started.

The tracked positions of body joints and other features are stored in a raw format, as provided by the system without any post-processing. This means that the three-dimensional recordings are currently in a different coordinate system than the matching two-dimensional versions, as described in the previous sections. Additionally, the gestures were not normalized to compensate for differences in the participants' height, or their position relative to the Kinect and camera. Because the Kinect has a relatively wide angle of view, and because OpenPose is likely to see human-like shapes in background objects, several recordings contained data for more than one person. Recordings for which this was the case were analyzed and any measurements not related to the participant performing the gesture were removed. Finally, all data were pseudonymized, and identifiable information was removed (e.g., email addresses from the log files).

## 5.5 Conclusion and discussion

In this chapter, we present a large dataset of iconic gesture recordings, collected in a naturalistic setting at a science museum and a music festival. Contrary to most existing gesture elicitation procedures, in our set-up participants were free to choose how they depicted a concept by gesturing, and they were distracted from the fact that they were being recorded. With this research we aim to contribute to the fields of gesture research and human-agent interaction in two ways. First, we provide a dataset that can be used as a basis for studies into human gesturing behavior — e.g., preferred modes of representation, differences based on age or culture, and changes in gesturing strategy after miscommunication occurs — showing the degree to which variation occurs in human-performed gestures. The dataset can be used for the design of an agent's capability to perform human-like gestures, and to recognize gestures performed by human interlocutors, taking into account this degree of variation. Second, we introduce the game of charades with a robot as a semi-structured elicitation procedure, which can be used to collect additional data in the future. To our knowledge this is the first publicly available elicitation method that employs a gameful interaction to collect gesture recordings.

This gameful elicitation method is able to bring gesture research out of the laboratory and closer to real-world settings. However, because the game restricted participants to only use their upper body, without support from speech, and because the other player was a robot that was not very good at recognizing the gestures, we imagine that the currently recorded gestures are more exaggerated (e.g., in terms of the size of the motions) than co-speech gestures used in everyday human-human conversation. It would be interesting to develop a variation of the system that is closer to the original game of charades, in which people are asked to describe an object, either using gestures or a combination of speech and gestures. In that case, the data would be less structured, because gestures no longer relate to specific cued objects but instead to object properties (e.g., 'big', 'heavy'), however this would result in more broadly usable gestures. It would also be interesting to record co-speech gestures during free-form conversation with a robot, and to see if people change their gesturing behavior when their conversational partner is a robot instead of another person. The current dataset, although it contains specific gestures for 35 concepts, can be used to study various aspects of general human gesturing behavior (e.g., repair strategies, variation in preferred modes of representation)[4]. In addition, these — arguably relatively expressive — gestures are useful in domains such as foreign language education, where it is important that their meaning is especially clear, even without speech. For example, we recently used a number of gestures from this dataset in an experimental study, in which a NAO robot was used as an English language tutor for children of 4–6 years old, and the gestures were implemented to support the children's learning process (Chapter 6).

It is important to stress that these data were collected in the field, and therefore will contain some degree of noise. There are examples where participants already started moving before the recording started, or where the system did not detect the end of the gesture properly and recorded additional movements that were no longer related to the gestures. These recordings were left as is on purpose, to give a realistic representation of the situations one could encounter when bringing this type of technology into the field, and to provide data that can be used to build solutions that can cope with these situations. As a concrete example, at one point during the experiment a participant was asked to perform a gesture for the concept *violin*, but instead showed a gesture that clearly referred to a *guitar*, another concept from our

[4] We used the current dataset after publication of this chapter in order to study, using computational analyses, whether gestures that are semantically related (e.g., *bird* and *airplane*) also have similar kinematics (Pouw et al., 2021). We found that it appears to be possible, at least to some extent, to derive semantic relatedness from gesture kinematics.

set. An additional research question that could therefore be answered using the dataset is how systems can be made intelligent enough to detect these discrepancies and handle them accordingly, for example by asking for clarification and performing the necessary relabeling autonomously.

To reduce the duration of the interaction we had to limit the number of concepts that each participant was asked to perform. Therefore, the dataset only contains recordings of 5 concepts per participant, instead of all 35. It is possible that the selection of concepts, and the order in which they were presented, has affected the resulting gestures. For example, both *car* and *bus* were included in the list. If participants were first presented with the cue for bus, they might only perform the act of driving, thinking that this was a unique enough description of the bus. However, if they had previously become aware that car was also included, they might have added an additional motion describing the shape of the bus, or the act of letting people board the bus, in addition to the driving motion to distinguish between the two related concepts.

The participants' preferred strategy for depicting the concepts using gestures may have further been affected by the images that were used as prompts. For example, the image for *bridge* (shown in Figure 5.2) contained a particular example with arches, which caused several participants to include an arching shape in their gesture. However, it is still unclear whether this priming effect shows for all of the included concepts. There could be concepts with a clear default mode of representation (Dargue & Sweller, 2018; Masson-Carro et al., 2017; Ortega & Özyürek, 2016; Ortega & Özyürek, 2020; van Nispen et al., 2014; van Nispen et al., 2017), which is then not affected by their representation in the images. This can be further investigated with the data we have available now, by measuring how often specific features from the images come up in the matching gestures.

There are several technical limitations to this method of data collection. The current version of the system relies on external devices — the Kinect and video camera — in order to record the gestures. We envision that in the future robots will have these features embedded, so that gesturing can become a more integral part of their abilities. This is a necessary step to make robots more inclusive by enabling them to communicate in situations where the effectiveness of spoken language is compromised, such as noisy environments or when the interlocutor has trouble understanding speech (e.g., due to being deaf or hard of hearing, or due to aphasia). In addition, the motion recording quality of the Kinect sensor is worse than that of a professional motion capture set-up. However, the portability of the Kinect, and the fact that it does not require any markers or special clothing made it more suitable to bring into a naturalistic setting such as the museum and music festival. We felt that this was also a more realistic representation of what robots of the near future would be able to do. Finally, we decided not to publish video data from the participants. Although this would have resulted in a higher level of detail, we thought that this would also increase the barrier for visitors to the museum and music festival to engage in the interaction, and might make those that did participate feel more aware of the fact that they were being recorded.

The recorded gestures were automatically mapped onto the robot, however the robot is more limited than humans in its ability to perform the gestures. As a result, it was often not clear to participants to which concept the robot-performed gestures belonged. We imagine that the performance of virtual agents or more articulate robots, both with more degrees of freedom, would be better. In future work we aim to extend the system to include support for these different agents. In addition, it might be possible to optimize the translation between the recorded gestures and the NAO robot specifically. In the aforementioned study in the field of education (Chapter 6), we applied a hybrid approach where we used recordings from the NEMO-Lowlands dataset as inspiration for the design of the gestures for a NAO robot, which were then recreated using key framing techniques (Chapter 6).

In this chapter we have only provided first explorations of the dataset. There are several aspects to the gestures that can be further quantified, pertaining to the chosen modes of representation, and to the way the motions were executed (e.g., size, complexity), both within and between different concepts. We expect these aspects to be influenced by factors such as age (Jain et al., 2016; Masson-Carro et al., 2015; Sekine et al., 2018; Stites & Özçalışkan, 2017), and whether this was a first or second attempt at performing the gesture. In future work we intend to perform a more in-depth and structured analysis of the data, in order to provide an overview of the degree of variation that exists within the set. This research can be further supported by the currently available software tools for (semi-)automatic gesture analysis (Pouw & Dixon, 2020; Trujillo et al., 2019).

In conclusion, we introduce a dataset of iconic gestures with a number of elements that set it apart from other currently available datasets: it includes a large number of recordings, from a diverse group of participants (e.g., children and adults), where participants were free to choose their gesturing method, and they were asked to perform a second attempt if the robot failed to recognize their first gesture, to provide insight into possible repair strategies that people use when non-verbal miscommunication occurs. Furthermore, the gestures were recorded by means of a semi-structured, gameful elicitation procedure. As a result, this dataset can be used for research into human gesturing behavior, and as input for various automated gesture analysis, recognition, and production algorithms. Finally, we have made the elicitation method publicly available, so that other researchers can extend the dataset in a consistent, structured manner.

* * *

*In this chapter, we introduced a dataset of human-performed gestures, which can be used to (computationally or manually) study human gesturing behavior, and as input for gesture production and recognition for social robots or virtual agents. Because the participants were given single word cues, they were free to choose what kind of gesture they wanted to perform. As a result, the dataset contains different gesture shapes for the same concept. Children and adults participated in this elicitation study, which enables us to investigate whether they choose different ways to represent concepts using gesture.*

*The next chapter describes a conceptual replication of our first study (Chapter 3), but now including animal names as well as words for which the gestures were less expressive (e.g., 'bridge'). The gestures are based on recordings from the dataset that was introduced in the current chapter.*

## 5.A Overview of three-dimensional recordings

Table 5.4: Number of three-dimensional recordings per concept.

| Concept | NEMO | Lowlands | Total |
|---|---:|---:|---:|
| Airplane | 74 | 25 | 99 |
| Bed | 80 | 32 | 112 |
| Bird | 72 | 27 | 99 |
| Boat | 74 | 29 | 103 |
| Book | 74 | 28 | 102 |
| Bridge | 78 | 32 | 110 |
| Bus | 76 | 30 | 106 |
| Car | 77 | 27 | 104 |
| Castle | 75 | 32 | 107 |
| Chair | 78 | 32 | 110 |
| Comb | 78 | 27 | 105 |
| Cow | 76 | 26 | 102 |
| Crocodile | 81 | 29 | 110 |
| Cup | 76 | 30 | 106 |
| Drum set | 75 | 34 | 109 |
| Fish | 83 | 25 | 108 |
| Guitar | 76 | 25 | 101 |
| Helicopter | 75 | 25 | 100 |
| Horse | 78 | 30 | 108 |
| Lamp | 71 | 26 | 97 |
| Motorcycle | 81 | 30 | 111 |
| Pencil | 85 | 31 | 116 |
| Piano | 78 | 29 | 107 |
| Pig | 75 | 29 | 104 |
| Scissors | 78 | 29 | 107 |
| Spoon | 80 | 29 | 109 |
| Stairs | 84 | 29 | 113 |
| Table | 79 | 31 | 110 |
| Toothbrush | 75 | 28 | 103 |
| Tortoise | 77 | 29 | 106 |
| Train | 74 | 25 | 99 |
| Triangle | 81 | 33 | 114 |
| Trumpet | 76 | 28 | 104 |
| Violin | 78 | 27 | 105 |
| Xylophone | 82 | 27 | 109 |
| Total | 2,710 | 1,005 | 3,715 |

## 5.B Overview of two-dimensional recordings

Table 5.5: Number of two-dimensional recordings per concept.

| Concept | NEMO | Lowlands | Total |
|---|---:|---:|---:|
| Airplane | 57 | 27 | 84 |
| Bed | 61 | 32 | 93 |
| Bird | 68 | 25 | 93 |
| Boat | 67 | 29 | 96 |
| Book | 62 | 30 | 92 |
| Bridge | 73 | 32 | 105 |
| Bus | 65 | 30 | 95 |
| Car | 74 | 27 | 101 |
| Castle | 66 | 31 | 97 |
| Chair | 63 | 30 | 93 |
| Comb | 59 | 27 | 86 |
| Cow | 65 | 20 | 85 |
| Crocodile | 75 | 29 | 104 |
| Cup | 65 | 27 | 92 |
| Drum set | 63 | 32 | 95 |
| Fish | 72 | 25 | 97 |
| Guitar | 72 | 21 | 93 |
| Helicopter | 60 | 23 | 83 |
| Horse | 70 | 30 | 100 |
| Lamp | 58 | 24 | 82 |
| Motorcycle | 69 | 28 | 97 |
| Pencil | 76 | 31 | 107 |
| Piano | 60 | 29 | 89 |
| Pig | 68 | 29 | 97 |
| Scissors | 58 | 27 | 85 |
| Spoon | 66 | 29 | 95 |
| Stairs | 77 | 26 | 103 |
| Table | 60 | 28 | 88 |
| Toothbrush | 65 | 28 | 93 |
| Tortoise | 65 | 29 | 94 |
| Train | 63 | 25 | 88 |
| Triangle | 71 | 33 | 104 |
| Trumpet | 58 | 28 | 86 |
| Violin | 62 | 26 | 88 |
| Xylophone | 64 | 25 | 89 |
| Total | 2,297 | 972 | 3,269 |

## Chapter 6

# Exploring the Benefits of Variation in Iconic Robot-Performed Gestures for Second Language Learning

*This chapter is based on:* de Wit, J., Brandse, A., Krahmer, E., & Vogt, P. (2020, March). Varied human-like gestures for social robots: Investigating the effects on children's engagement and language learning. In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (pp. 359-367). *Open practices:* The materials are available at https://github.com/l2tor/animalexperiment/tree/variation. The study and analysis plan were preregistered on AsPredicted: https://aspredicted.org/wj24k.pdf.

## Abstract

To investigate whether a humanoid robot's use of gestures improves children's learning of second language vocabulary, and if variation in gestures strengthens this effect, we conducted a field study where a total of 94 children (aged 4–6 years old) played a language learning game with a NAO robot. The robot either used no gestures at all, repeated the same gesture every time a target word was presented, or produced a different gesture for each occurrence of a target word. We found that, contrary to what the majority of existing research suggests, the robot's use of gestures did not result in increased learning outcomes, compared to a robot that did not use gestures. However, engagement between child and robot was higher in both the repeated and varied gesture conditions, compared to the condition without gestures. An exploratory analysis showed that age played a role: the older children in the study learned more than the younger children when the robot used gestures. It is therefore important to carefully consider the design and application of robot gestures to support the learning process. The contribution of this work is twofold: it is a conceptual reproduction of a previous study, and we have taken first steps toward exploring the role of variation in gestures. The study was preregistered, and all materials are made publicly available.

## 6.1 Introduction

Manual gestures (Kendon, 2004) are an essential part of our everyday communication with other people: we produce them naturally to support our thinking process, and use them to avoid miscommunication (Hostetter, 2011). Specifically iconic gestures — a subset of gestures where the movements are meaningfully linked to the concept that is referred to (McNeill, 1992) — are known to be a valuable support mechanism in education, resulting in improved learning outcomes and higher levels of engagement from the student with the educational process (Kelly et al., 2008; Valenzeno et al., 2003; Wakefield et al., 2018). The present work focuses on the domain of second language (L2) learning, where gestures have been shown to contribute to increased vocabulary acquisition (Hald et al., 2016; Macedonia et al., 2011; Tellier, 2008). They enable "grounding" of new knowledge in existing sensorimotor experiences (Barsalou, 2008). For example, when teaching students about the word *ball* in a second language, by accompanying this unknown word with an iconic gesture depicting the underlying concept of a ball (e.g., by molding its shape or by bouncing an imaginary ball) we provide additional scaffolding to create a link between the new form (the L2 word) and the learner's existing knowledge of its corresponding meaning.

With an increasing research interest into using robots in contexts where they are expected to interact socially with humans, such as education (Belpaeme et al., 2018) and specifically second language learning (Kanero et al., 2018b; van den Berghe et al., 2019), a number of groups have started exploring whether gestures result in similar positive effects when they are being performed by a robot instead of a human. A survey comparing robots to virtual agents indicates the robot's ability to move and perform gestures in the physical world to be one of its key advantages over screen-based alternatives (Li, 2015). Observed effects of a robot's use of gestures include increased memorization of story details by the listener (Bremner et al., 2011; Huang & Mutlu, 2013), better human-robot collaborative task performance (Breazeal et al., 2005), and higher levels of engagement with the robot (Bremner et al., 2011; Gielniak & Thomaz, 2012; Sidner et al., 2005). Furthermore, a robot that gestures is generally perceived more positively (Aly & Tapus, 2013; Asselborn et al., 2017; Salem et al., 2013a), especially when its motions are exaggerated and cartoon-like (Gielniak & Thomaz, 2012).

However, applications of robots that gesture in an educational context, and specifically in (second) language learning, remain underresearched. One example can be found in first language learning with adults, where participants that interacted with a robot that used iconic gestures had better learning outcomes than those that did not receive support from the robot's gestures (van Dijk et al., 2013). We recently conducted two studies in second language learning, both with young children as participants, with mixed results. A first exploration showed increased vocabulary retention over time, as well as higher levels of engagement with the robot for children that received additional support from the robot in the form of iconic gestures compared to children that were not presented with gestures (Chapter 3). Our second study, although similar in design, did not see such an effect on learning outcomes (Vogt et al., 2019). The two studies differed in their duration (number of sessions) and the vocabulary that was taught. The first study consisted of only one session, and taught six animal names while the second study was longitudinal (seven sessions) and contained a larger set of more abstract target words (e.g., *more* and *above*), potentially leading to a lower degree of iconicity in the gestures. In view of these conflicting findings, we set out to test the effects of gestures in a single lesson of second language learning with a robot, but using a more diverse set of target words. This part of the current study is a conceptual replication (Zwaan et al., 2018) of our previous work (Chapter 3) as it includes two of the same experimental conditions from the original study — one where the robot uses iconic gestures, and one where the robot does not use any gestures — although with different target words and several improvements to the measurements. It is important to highlight that the design of robot gestures in earlier studies often only relied on the intuitions of the researchers. Here, instead of defining and designing gestures ourselves, we looked at existing sources to see how humans depict the words. We expected to find results that match those found in the original study, leading to the following two hypotheses:

H1 Children will learn more target words in a second language (H1a) and remember them better (H1b) when a robot produces iconic gestures for the target words than when the robot does not produce such gestures.

H2 Children are more engaged when interacting with a robot that produces iconic gestures for the target words than with a robot that does not produce such gestures.

When we gesture, we choose which aspect of a concept to describe with our movements, and which strategy — or mode of representation (Müller, 2014) — we use to depict it: do we focus on shape, such as the roundness of a ball, or rather the act of throwing or shooting a ball? Although people generally have default strategies, there is still a degree of individual variation in how we produce gestures (Masson-Carro et al., 2017; Ortega & Özyürek, 2016; van Nispen et al., 2014; van Nispen et al., 2017). This variation can partly be explained by cultural differences (Kita, 2009), as well as age (Masson-Carro et al., 2015; Sekine et al., 2018; Stites & Özçalışkan, 2017). Children tend to maintain a smaller symbolic distance to the concept they are describing, which means they will often use a larger gesture space (body parts or the full body), while adults tend to take the "outsider looking in" perspective, and use only their hands to represent objects or characters (Sekine et al., 2018). For example, when depicting a *pencil* children are more likely to raise their hands above their head in a pointy shape, representing the pencil with their entire body, compared to adults who generally use their hands to display the act of writing, or outline the shape of a pencil.

How we represent concepts in gesture might also be related to what Piaget and Cook (1952) defined as schemata, mental representations describing the objects and concepts we know, and any past experiences or actions related to these objects. For example, our schema of a toothbrush could include some of its typical visual features, as well as the act of brushing our teeth. As we develop and experience more aspects of a particular concept, our schema of this concept becomes more elaborate. A related framework is variation theory, which states that the object of learning (e.g., in our case the concepts to which we want to link L2 words) may be perceived differently between people, where one learner might focus on different aspects than another (Marton & Booth, 2013). This theory suggests to add variation to learning examples, thus highlighting multiple features of the object of learning. Both theories identify a certain amount of pre-existing knowledge in the learner — which varies between individuals, and grows with experience — to which new features can be added (Hanfstingl et al., 2019). This, combined with the fact that we use different strategies for producing gestures, raises the question whether we also have different preferences and skills when it comes to understanding and integrating gestures.

There appears to be no existing research that looked into possible benefits of using variation in gestures to support learning. However, there have been studies in the context of second language learning where variations were introduced in the number of different speakers (Barcroft & Sommers, 2005), reporting better learning outcomes compared to the use of a single speaker. Another study varied the images that were used to support second language learning (Sommers & Barcroft, 2013). Contrary to what was found with variations in speech, this had an adverse effect on the number of newly acquired vocabulary items compared to repeating the same image. The researchers suggest that this may have been caused by shifting the focus from the form (the L2 word and how it is pronounced — the new knowledge that is being taught) to the existing meaning assigned to it by the learner (represented by the image). In the present study we have kept both speech and supporting imagery constant throughout the interaction, while variation is added to the additional gesture modality.

Based on the aforementioned theories, we hypothesize that variation in the robot's gestures results in a greater chance that the gestures align with existing salient features of the underlying concept that are already part of the learner's schemata. Furthermore, by presenting several different features the learner might create a stronger link between the word and the underlying concept, rather than merely linking words to specific stimuli. Existing research also indicates that children are more engaged when interacting with robots that show less repetitive behavior (Tanaka et al., 2007). We therefore hypothesize:

H3 Children will learn more target words in a second language (H3a) and remember them better (H3b) when a robot produces a different iconic gesture every time a particular target word is presented than when the robot produces the same iconic gesture every time a target word is presented.

H4 Children are more engaged when interacting with a robot that produces a different iconic gesture every time a particular target word is presented than with a robot that produces the same iconic gesture every time a target word is presented.

Because the ability to interpret gestures grows with age (Novack et al., 2015; Stanfield et al., 2014), we also explore whether differences in age within our participant group have affected their learning outcomes or engagement. The present study adds to existing research in the field of human-robot interaction and gesture studies by verifying whether the previously observed positive effects of gestures persist when the concepts that are taught are more diverse. Furthermore, we investigate whether the previously unresearched addition of variation in a robot's repertoire of gestures further increases these effects. We also propose several improvements to the process of measuring learning outcomes and engagement, with the goal of improving the reliability of our findings. Our hypotheses and planned statistical analyses were preregistered[1], and all of the source code and materials needed to replicate this study are made publicly available[2].

## 6.2 Design of the interaction

We used the one-on-one tutoring interaction from our previous study (Chapter 3), in which a child and a SoftBank Robotics NAO robot together played a simplified version of the game *I spy with my little eye*, which is described in more detail below. Two minor changes were made to the original source code. First, the target words were changed to include a more diverse set of objects: bridge, horse, pencil, spoon, stairs, and turtle. Second, we implemented the additional experimental condition in which the robot used a different gesture every time a target word was presented. The five available gestures for each concept were randomized for each participant, so that no order effects could occur. We now briefly explain the process of designing and validating the gestures, and the workings of the educational game that was used.

### 6.2.1 Gestures

In order to ensure that only gestures that participants were likely to recognize were used, all of the robot's depictions were based on an existing dataset of recordings from humans producing silent gestures in the context of a game of charades with a robot (Chapter 5). We based our choice of target words on the availability of varied examples within this dataset, while ensuring that they covered a diverse range of categories (e.g., tools, static objects, animate objects). We also took into account the age of acquisition (Kuperman et al., 2012) for the words, so that the children in our study should know them in their first language. Although the dataset includes three-dimensional Kinect recordings, directly mapping those onto the NAO robot resulted in noisy and unclear gestures. We therefore recreated them by defining key frames using the Choregraphe software that is distributed with the NAO robot (Pot et al., 2009), while staying true to the recorded gestures as much as possible. This is a common workflow for creating robot motion that was also used in the original study (Chapter 3), but now based on examples of people performing the gestures rather than the researchers' frame of reference. Out of the 30 gestures that were

[1] https://aspredicted.org/wj24k.pdf

[2] https://github.com/l2tor/animalexperiment/tree/variation

implemented, 16 were based on recordings from male performers and 14 from females. Nineteen gestures were recorded from primary school-aged children (6–12 years old), another 10 by adults (20–62 years old), and 1 by a teenager (15 years old).

After recreating the gestures, we video recorded the robot as it performed them and evaluated their clarity by means of an online questionnaire. A total of 19 participants (10 male and 9 female, $M_{age} = 38$ years, $SD = 15$ years) was recruited through convenience sampling. They were shown a video of a gesture and were asked to select the matching concept out of all six included in the study, to investigate whether the gestures were unique enough within the set of six target words. Out of the 30 gestures, 8 scored poorly (< 60% accuracy), 9 scored moderately (60–70%), and 13 scored strongly (> 70%). Based on these findings and additional qualitative feedback, 14 of the gestures were revised to more closely match the human-performed examples from the dataset. Figure 6.1 shows the five variations for the target word *turtle*. For the experimental condition where the robot did not vary its gestures, we implemented the example that scored highest in the questionnaire (the middle image in Figure 6.1 for *turtle*).

### 6.2.2 Language learning game

To train the six target words in the L2 (English), the child and the robot engaged in a simplified version of the game *I spy with my little eye*. The set-up of the experiment included the robot, and a tablet on which the child was able to select answers (see Figure 6.2). During the training the child sat at a table on which the tablet was placed at a slightly tilted angle. The robot was standing opposite the child and was put in breathing mode, meaning that it moved its head and arms around slightly and

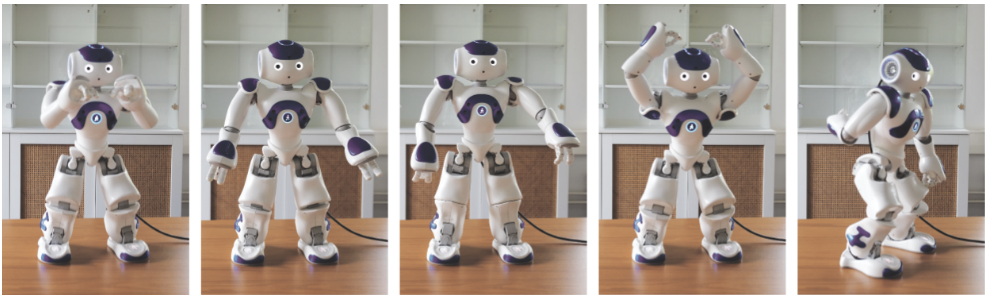


Figure 6.1: We investigated whether a robot can use iconic gestures to support its teaching activities, and if it helps to add variation to these gestures. These are the gesture variations for *turtle*. Videos available at https://tiu.nu/hri20-gestures.

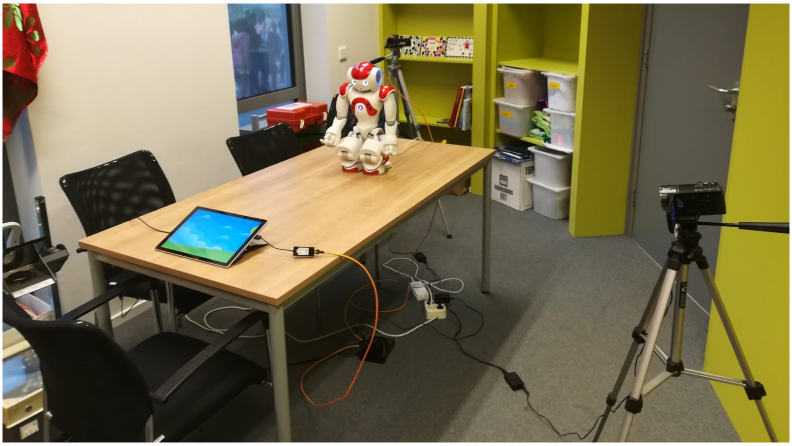


Figure 6.2: The set-up of the experiment at one of the schools.

shifted its weight between its legs in order to appear more lifelike.

The robot started by greeting the child with his/her name and then explaining the game, after which the child was asked to indicate whether he or she understood the instructions by touching either a green or red smiley face on the tablet. If the child did not understand the concept of the game, a researcher stepped in to provide further explanation. The game then started with two practice rounds, which were always for the target word *horse* — one in the first language, or L1, Dutch and one in the L2, English — followed by 30 rounds of the game. Each round started with the robot calling out a target word: "Ik zie ik zie wat jij niet ziet, en het is een... *horse*" ("I spy with my little eye a... *horse*"). Three images then appeared on the tablet screen: the correct answer, along with two randomly chosen distractor images (Figure 6.3). Three images were shown to ensure that the difficulty level while children were still learning was lower than during the post-tests (with six images). The robot provided feedback in response to the child's answer, in which the L2 target word was mentioned again but without any gestures. If the child selected the wrong image, a "repair round" took place where the robot spied the same word once more, but now only the correct image and one distractor image — the previously given answer — were shown.

During the 30 rounds, each of the six target words was presented five times in total, but their order was randomized. In the experimental condition with repeated gestures, the same gesture was used for all five times each target word was presented. In the condition with variation in gestures, the target word in every round was

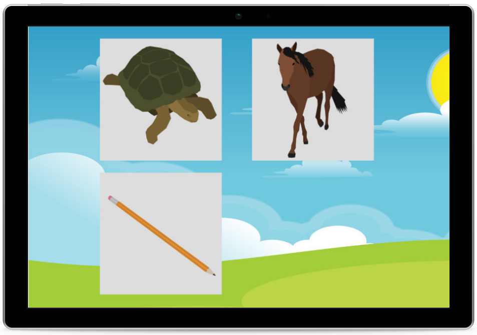


Figure 6.3: Children provided answers on a tablet screen.

accompanied by a different gesture for that word, but for repair rounds the same gesture from the main round was used. The condition without gestures was identical to the others, but no gestures were used at all. After finishing all 30 rounds, the robot said goodbye to the child. The researcher had a control panel where the child's name was entered, which was used by the robot to personalize the introduction. After pressing a *Start* button, the robot operated fully autonomously, but the interaction could be paused at any time by the researcher if a break was needed. Autonomous behavior was possible by minimizing the complexity of the interactions — the robot did not "listen" to the child, answers to its questions were given through the tablet device.

## 6.3 Methodology

In order to investigate whether the robot's use of iconic gestures resulted in increased learning outcomes and higher levels of learner engagement compared to a robot that does not use such gestures, and to see whether variation in gestures increases learning outcomes and engagement more than repeating the same gesture, we conducted an experiment with the following three experimental conditions: (1) No gestures, where no iconic gestures were included at all; (2) Repeated gestures, where the robot used the same gesture every time a target word came up in the game; (3) Varied gestures, where the robot used five different gestures — a new one for every time a target word came up in the game. Other than these differences in the robot's use of gestures, the experimental conditions were identical, and all children engaged in the same previously described language learning game.

Table 6.1: Demographic information of study participants.

| Experimental condition | N | Age (Y;M) ± SD (M) | Boys/girls |
|---|---|---|---|
| No gestures | 33 | 5;3 ±9 | 51% / 49% |
| Repeated gestures | 32 | 5;2 ±9 | 56% / 44% |
| Varied gestures | 29 | 5;4 ±8 | 41% / 59% |
| Total | 94 | 5;3 ±9 | 50% / 50% |

### 6.3.1 Participants

A total number of 116 children, recruited from two different primary schools in the Netherlands participated in the study. However, 22 participants had to be excluded due to technical or procedural issues ($N = 12$), bilingualism ($N = 3$), English pre-test scores that were too high (more than four out of six correct, $N = 3$), and missing results due to drop-out ($N = 4$). As a result, the data of 94 children were included in our analyses. The participants were pseudo-randomly assigned to one of the three conditions with a balanced distribution of age and gender (see Table 6.1 for demographic information). The study was approved by the research ethics committee of Tilburg University. Informed consent was given by the parents of the children prior to their participation.

### 6.3.2 Pre-test and post-tests

Children's vocabulary knowledge was measured at different times by means of a test, where images for all six target words were presented on a laptop screen (Figure 6.4). A voice recording then asked the child to identify the matching image for a particular target word: "Waar zie je een... [word]?" ("Where do you see a... [word]?"). To reduce bias due to random guessing, in the L2 version each target word was tested three times, yielding a total of 18 test items. To ensure that the test also measured generalizable knowledge, such that the L2 words were not simply linked to the images as they came up in the training session with the robot but rather to the underlying concepts, each of the three times a different image was used: either the same image from training, a photorealistic version, or a line drawing. A target word was scored as correct if the child managed to identify it correctly in at least two out of the three rounds, resulting in a final score of 0–6. In the L1 version of the test each target word was tested only once to save time, and because we assumed that children already knew all of the words in their first language.

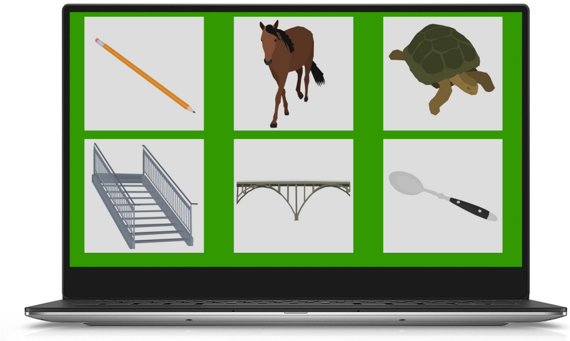


Figure 6.4: “Game” used to test children’s word knowledge.

### 6.3.3 Procedure

#### Group introduction

Based on our previous experience working with children and robots, as well as reports from other studies (Fridin, 2014; Vogt et al., 2017a), we organized a group introduction to help the children feel at ease with the robot. This was done for entire classrooms at the same time, with the teacher also present. In this session the researchers introduced the robot and demonstrated some of its features. Children were then allowed to shake hands with the robot and put it to bed. The introduction took approximately 15 minutes.

#### Pre-test

To measure the pre-existing knowledge of the target words in the L1 and L2, each child was retrieved from the classroom and was asked to complete the test on the laptop, as previously described in Section 6.3.2. The pre-tests were planned on the same day as the group introduction or shortly thereafter, without the robot present. The tests took approximately 10 minutes and included additional questions related to the children’s perception of the robot which are not further analysed here (and were not part of the preregistered analyses).

#### Training and immediate post-test

The actual training session was scheduled at least one day after the pre-test. The child was retrieved from the classroom and brought to the experiment room. This session consisted of three parts. First, the child was invited to complete a short “game” on the laptop, where each of the six target words was introduced three times (“Look, this is a [word]. Do you see the [word]? Click on the [word].”), while the corresponding image was shown on the screen. This was done to familiarize the child with the target words, so that they had some prior knowledge before practicing with the robot. The child was then invited to go sit at the table with the tablet and robot, and play the game of *I spy with my little eye* for 30 rounds as described previously. After completing the interaction with the robot, children were asked once more to sit down at the laptop and complete the English post-test. The total duration of this session was 25–45 minutes, depending on experimental condition — gestures slowed down the training — and on the number of repair rounds needed. The researcher was always present during the session, although he or she was instructed to act busy to avoid having the child turn to them for task-related feedback.

#### Delayed post-test

Between one and two weeks after the training session with the robot, each child was retrieved from the classroom once more for a delayed post-test. This test was identical to the immediate post-test administered after the child's interaction with the robot, and lasted approximately three minutes.

### 6.3.4 Analyses

In line with the preregistration and with the original study, we have conducted a series of ANOVAs with difference scores between the post-tests and pre-test. However, after submitting the preregistration we realized that a single mixed ANOVA would be more optimal, since it reduces the risk of type I errors by minimizing the amount of statistical analyses required. For consistency, we present the results of both analyses. Engagement was annotated by extracting two video clips from each child's interaction with the robot, one from the 4th and one from the 24th round of training. Each clip lasted two minutes and was annotated for two different measures of engagement: task engagement and social engagement with the robot. The ratings were based on a coding scheme that was recently developed[3], which resulted in a score for each type of engagement on a nine-point scale (1–9). Note that engagement is considered as a measure of how actively the child was involved with the robot or the task, not whether this was positive (constructive) or negative (destructive) involvement. The Pearson correlation between task and robot engagement was .60 ($p < .001$).

In comparison to our previous analysis of engagement (Chapter 3) we aimed to improve robustness by increasing the length of each clip (two minutes rather than

[3] https://github.com/l2tor/codingscheme

five seconds), by rating engagement across two distinct dimensions rather than a single all-encompassing measurement, and by using coding schemes upon which to base these ratings. Instead of distributing an online questionnaire, the ratings were now performed by one of the researchers. To test the reliability of our measures, 50 video clips (taken from 25 different sessions) were annotated by a second rater who did not participate in the original data collection and was not familiar with the specifics of the experimental conditions. The intraclass correlation (ICC) estimates and their 95% confidence intervals were calculated using SPSS version 24 based on a single rater, consistency, two-way random effects model. This resulted in a 95% CI of [.45, .78] for task engagement (considered poor–good, cf. Koo & Li, 2016), and a 95% CI of [.55, .83] for robot engagement (moderate–good). Based on this ICC we proceeded with the ratings of a single rater in our analyses.

## 6.4 Results

### 6.4.1 Preregistered analyses

**Learning outcomes**

Figure 6.5 shows the mean scores on the three tests per condition, indicating a similar increase in vocabulary knowledge over time between conditions.

A 3 (experimental condition) × 3 (test time) mixed ANOVA was used to evaluate children's learning outcomes, with scores on the test tasks (0–6) as dependent variable, experimental condition as between-subjects independent variable, and time (pre-test, immediate post-test, and delayed post-test) as within-subjects independent variable. The analysis showed a significant effect of time, $F(2, 182) = 45.70, p < .001$, $\eta_p^2 = .33$, indicating that children learned L2 vocabulary from their interactions with the robot regardless of condition. Pairwise comparisons using Bonferroni correction show a significant difference between the immediate post-test and the pre-test, $M_{dif} = 1.10, p < .001$, and between the delayed post-test and the pre-test, $M_{dif} = 1.41, p < .001$. However, there was no significant difference between the delayed post-test and the immediate post-test, $M_{dif} = 0.30, p = .09$. There was no main effect of condition, $F(2, 91) = 0.38, p = .68$, and no significant interaction between experimental condition and time, $F(4, 182) = 1.58, p = .18$, indicating that the robot's use of gestures — either repeated or varied — did not affect learning outcomes[4].

[4] For consistency with the preregistration and the analyses in the original study, we also performed a combination of t-tests and separate ANOVAs on difference scores. The results are identical to the

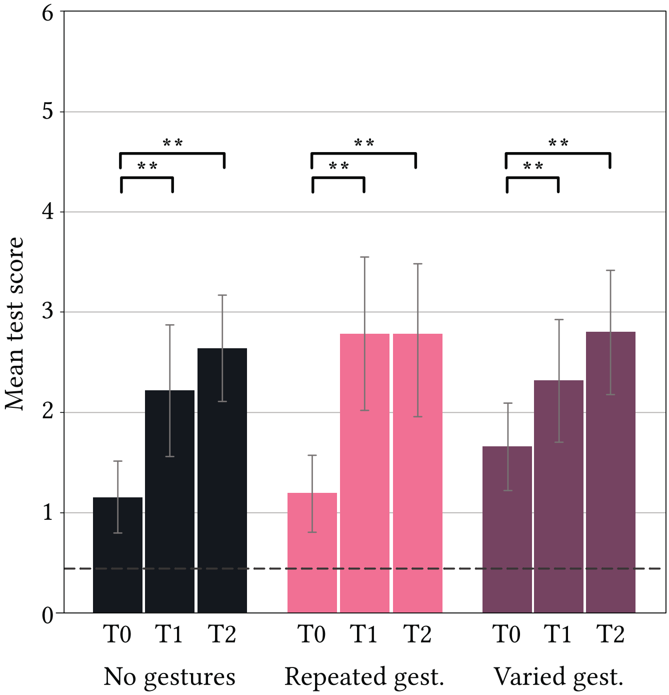


Figure 6.5: Mean test scores as a function of experimental condition (** $p < .001$). T0 = pre-test, T1 = immediate post-test, T2 = delayed post-test. Chance level (horizontal line) was 0.44. The error bars are the 95% CI.

## Engagement

Figure 6.6 visualizes task engagement (left) and social engagement with the robot (right), measured at rounds 4 and 24. A clear drop between rounds 4 and 24 can be observed for both types of engagement. Although task engagement levels are similar between conditions, children in the experimental condition without gestures are less engaged with the robot than those in both gesture conditions.

To evaluate whether the robot's use of gestures affected children's engagement, we conducted a 3 (experimental condition) × 2 (time) mixed MANOVA with the task and robot engagement ratings as dependent variables, time (round 4 and round 24) as within-subjects independent variable and experimental condition as between-subjects independent variable. This shows a significant effect of time, Wilk's $\Lambda = .30, F(2, 90) = 107.76, p < .001, \eta_p^2 = .71$, indicating a drop in engagement between

mixed ANOVA approach (a significant effect of time but not condition), with the exception of the difference between the delayed post-test and immediate post-test scores, which now also reached significance.

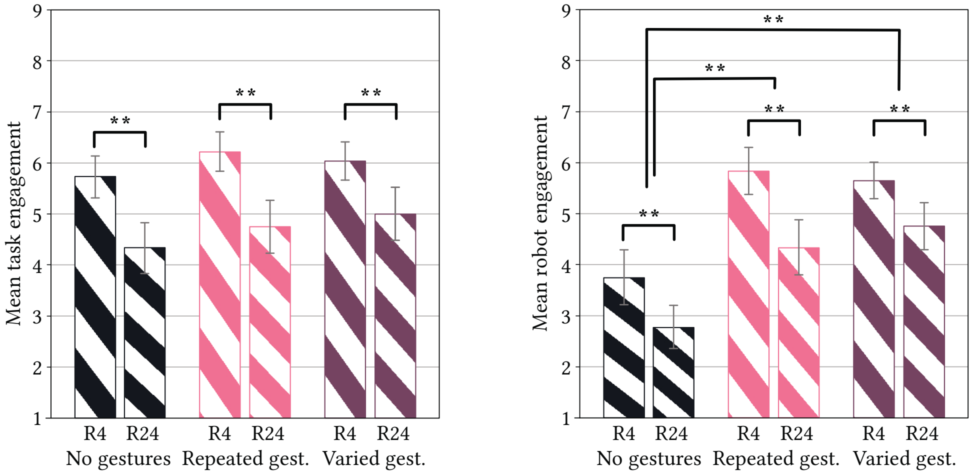


Figure 6.6: Task (left) and robot (right) engagement ratings for rounds 4 and 24, by condition (** $p < .001$). The error bars are the 95% CI.

rounds 4 and 24. This effect was found for task engagement, $F(1, 91) = 132.26, p < .001, \eta_p^2 = .59$, and for robot engagement, $F(1, 91) = 134.79, p < .001, \eta_p^2 = .60$.

The analysis also showed a main effect of experimental condition, Wilk's $\Lambda = .60, F(4, 180) = 13.20, p < .001, \eta_p^2 = .23$, indicating differences in average engagement throughout the interaction. This difference was only significant for robot engagement, $F(2, 91) = 25.9, p < .001, \eta_p^2 = .36$, and not for task engagement, $F(2, 91) = 1.88, p = .16$. A post-hoc analysis using Bonferroni correction showed that average robot engagement was significantly higher in the repeated gestures condition ($M_{dif} = 1.82, p < .001$), as well as in the varied gestures condition ($M_{dif} = 1.93, p < .001$), compared to the condition without gestures. The difference between the varied and repeated gesture conditions was not significant ($M_{dif} = 0.06, p = 1.0$). The interaction between time and condition was not significant, Wilk's $\Lambda = .90, F(4, 180) = 2.32, p = .06$, showing no effect of the robot's use of gestures on the change in engagement over time.

### 6.4.2 Exploratory Analysis of Age

Existing literature indicates that our ability to recognize and understand gestures grows with age (Novack et al., 2015; Stanfield et al., 2014). Additionally, we intuitively observed variations in how children of different ages interacted with the robot. Figure 6.7 shows a linear fit to children's difference scores on the immediate (left) and delayed (right) post-tests, indicating that age affected children's performance, especially in both experimental conditions where the robot used gestures. We ran the same mixed ANOVA with test scores as dependent variable, and time and condition as independent variables, now adding children's age in months at the time of the experiment as a covariate. This showed a significant main effect of age, $F(1, 90) = 19.30, p < .001, \eta_p^2 = .18$. The interaction between age and time was also significant, $F(2, 180) = 10.59, p < .001, \eta_p^2 = .11$, indicating that older children that participated in the study learned significantly more from the interaction than younger children. To further explore whether this effect of age was influenced by the robot's use of gestures, we split our data by experimental condition and ran the same analysis. This showed a significant interaction effect of age and time within the repeated gestures condition, $F(2, 60) = 7.83, p = .001, \eta_p^2 = .21$, and within the varied gestures condition, $F(2, 54) = 7.87, p = .001, \eta_p^2 = .23$, but not within the condition without gestures, $F(2, 62) = 0.74, p = .48$.

To investigate whether age also influenced children's levels of engagement, we ran the previously described mixed MANOVA with both measures of engagement as dependent variables, adding age as a covariate. This showed a main effect of age, Wilk's $\Lambda = .91, F(2, 89) = 4.41, p = .02, \eta_p^2 = .09$. This effect was only significant for task engagement, $F(1, 90) = 5.29, p = .02, \eta_p^2 = .06$, where the older children in the experiment showed higher task engagement than the younger children. There was no main effect for robot engagement, $F(1, 90) = .002, p = .97$, and no significant interaction effect between age and time, Wilk's $\Lambda = .97, F(2, 89) = 1.18, p = .31$.

## 6.5 Discussion

This chapter describes a study that investigated the potential benefits of a robot's use of gestures in second language tutoring. We compared between a robot that repeated the same gesture for each concept, one that varied its gesture repertoire, and one that did not use gestures at all, and we measured how this affected children's learning outcomes and engagement with the task and with the robot. The contribution of this work is twofold. First, it is a conceptual replication of a previous study (Chapter 3) with a shift toward a more diverse set of target words. Our goal was to verify whether our previous findings persist, especially in light of conflicting findings regarding robot-performed gestures in other studies (e.g., Vogt et al., 2019; Chapter 4). Several steps have been taken to improve the reliability and reproducibility of the study. These include various changes to the measures such as testing each target word multiple times, and the use of a coding scheme for rating children's

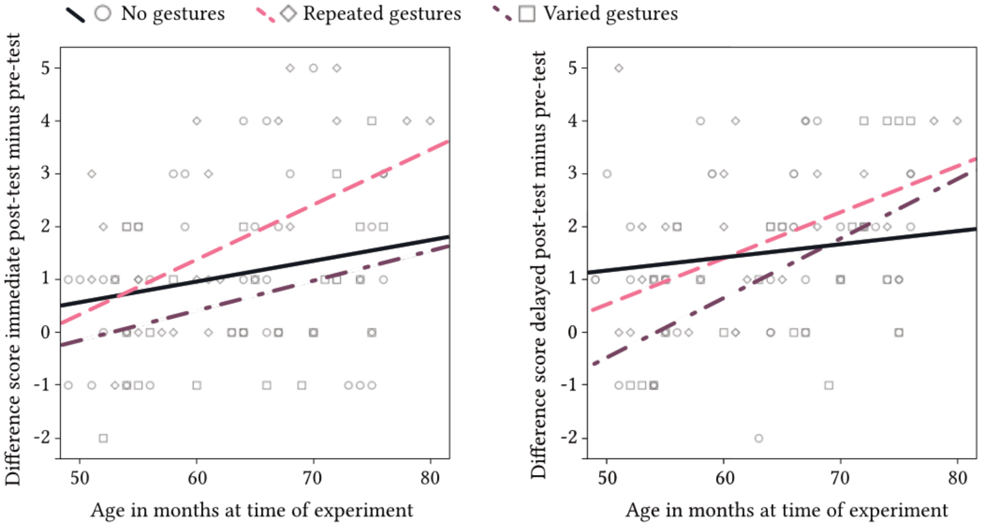


Figure 6.7: Linear fit to the difference scores on the immediate (left) and delayed (right) post-tests compared to the pre-test per condition, relative to children's age.

engagement. Second, despite the assumed importance of variation for educational purposes (Marton & Booth, 2013; Piaget & Cook, 1952) we did not find any existing research in this direction. Therefore we added an experimental condition where the robot introduced variation by performing different gestures for each concept. Our results show that a single tutoring session with the robot helped children acquire new L2 vocabulary, and retain this knowledge over time. Children on average learned 1.10 new words on the immediate post-test, and 1.41 on the delayed post-test — similar results to those in the original study (Chapter 3). This may not seem like a substantial increase, however these were young children and the results were obtained after a single training session of approximately 15 minutes. Other word learning studies with robots have shown similar results (Belpaeme et al., 2018; van den Berghe et al., 2019).

Contrary to the original study we did not find support for our first hypothesis that children would learn and remember more words when the robot used gestures than when the robot did not use gestures. This could be caused by the fact that we introduced more diverse and potentially more complex target words in the current work compared to the animal names in the original study, with perhaps less iconic gestures as a result. Because the overall number of words learned is similar across both studies, we can assume that the English words themselves were not necessarily more difficult to learn. The difference therefore appears to be in the gestures, where children found it harder to understand the gestures in the current study. It would be interesting to further investigate which exact characteristics of the gestures are responsible for these difficulties with their interpretation.

Older children in our study did appear to understand and benefit from the robot's gestures, while younger children did not. Although literature indicates that children learn how to make sense of iconic gestures at a slightly younger age than the age of participants in our study (Novack et al., 2015; Stanfield et al., 2014), the ability to interpret gestures could be reduced when the interaction involves a robot instead of a human, and when it is mediated by a tablet device. The robot's gestures appear to have a detrimental effect when they are not understood, which may have been caused by distraction, confusion, and the additional cognitive load from attempts to observe and make sense of these gestures. These findings underline the importance of properly designing the robot's gestures. Previous research often included gestures that were designed by the researchers, but in this work we based the design on a dataset with recordings of mostly children performing gestures (Chapter 5). The clarity of the robot's gestures was evaluated with 19 judges, and the consistency of the ratings showed that this sample size was sufficient. However, the process of designing gestures could be further improved in two ways. First, it would be better to evaluate the gestures with children from the same age group that participated in our study instead of adults. However, we believe that a task to judge the meaning of gestures is difficult for children this young, so this should perhaps be done in the form of a guessing game. Second, based on the ratings we made several improvements to the gestures, but these were not evaluated. We are confident that these changes resulted in better gestures since they now align more with the original human-performed examples, but in future work we would take a more iterative approach and conduct multiple evaluations.

Our second hypothesis stated that children would be more engaged with a robot that produces iconic gestures, than with one that does not produce gestures. This hypothesis finds partial support in a higher average robot engagement, however no significant effects on task engagement are found. These findings are consistent with literature on the effects of robot gestures on engagement (Bremner et al., 2011; Gielniak and Thomaz, 2012; Sidner et al., 2005; Chapter 3). We conjecture that the main reason for higher robot engagement is that the robot displayed more bodily movements in the gesture conditions, which can cause the robot to be perceived as more friendly and human-like (Asselborn et al., 2017), resulting in a higher level of engagement with the robot as children enjoyed the interaction more. Engagement with the task was influenced by age, however this does not seem to relate to the robot's use of gestures.

By introducing variation in the robot's gestures, and thereby highlighting different features of the object of learning (cf. Marton & Booth, 2013), we aimed to provide greater support to the learning process compared to using repeated gestures. We also expected this variation in the robot's behavior to further increase children's engagement with the robot (cf. Tanaka et al., 2007). However, we did not find support for hypotheses H3 and H4 which stated that the robot's use of varied gestures would lead to better learning outcomes and higher levels of engagement than repeated gestures. This does not align with existing findings in literature regarding positive effects of speaker variation (Barcroft & Sommers, 2005), nor detrimental effects of image variation (Sommers & Barcroft, 2013). Moreover, with multiple gestures for the same concept it is more difficult to measure what the contribution of each individual gesture was to children's learning outcomes and engagement. We believe more research is needed to further investigate possible differences between variation and repetition of gestures. The current study consisted of a single tutoring session and therefore did not investigate any potential long-term effects that variation in gestures might have. Furthermore, different results could be observed for older children or adults, and the use of varied gestures could have affected other factors that were not measured in the current study, such as perception of the robot (e.g., human-likeness, intelligence, character) or overall enjoyment. With younger participants it remains a challenge to investigate these aspects of a robot's appearance and behavior.

## 6.6 Conclusion

This chapter documents a study that was conducted to investigate whether a robot's use of iconic gestures affects learning outcomes and learners' engagement. Furthermore, a robot that varied its gesture repertoire for a particular concept was compared with one that always repeated the same gesture. The results of the study show that there are advantages to having a robot perform gestures when teaching children L2 vocabulary, in the form of higher engagement and — for the older children in the study — increased learning gain, although no additional benefits were found for varied gestures. Based on existing literature into robot-performed gestures (e.g., Bremner et al., 2011; Huang & Mutlu, 2013; Li, 2015; van Dijk et al., 2013) we have reason to believe that our findings generalize to different target groups, educational domains, and robotic platforms, and we imagine that robots in the future will become capable of performing increasingly more human-like motions. The design of the interaction, the gestures, and the study itself are documented in this chapter to serve as a basis for future research. We envision two main avenues for future work: (1) the design of the robot's gestures, and how this affects their comprehensibility for different ages, and (2) a further exploration of variation in gestures: does it have different effects on older learners, and does it change the way the robot and the interaction are perceived?

∗ ∗ ∗

*In the current chapter, we further investigated the mixed results found in the two previous studies (Chapters 3 and 4). We took the relatively simple game that was used in the first study, but included concepts with less expressive gestures, similarly to the second study. The results showed no significant effect of the robot's use of iconic gestures on learning outcomes. However, older children in the study again benefited more from the robot's use of iconic gestures than younger children did. These results further support the theory that within the age range of our participants (4–6 years old), there appears to be a point at which children learn to (better) make use of the robot's iconic gestures. In addition, children on average showed higher levels of engagement with the robot, but not with the task, when the robot used iconic gestures, compared to when it did not perform gestures.*

*This chapter also serves as a first exploration of introducing variation in the robot's gesturing behavior. We designed five different gestures for each concept, that were pseudorandomly performed by the robot. Introducing variation did not appear to have any benefits, but there were also no drawbacks, in terms of learning outcomes and levels of engagement, compared to having only one gesture for each concept. More research is needed to investigate the effects of variation on long-term engagement, and on the way the robot is perceived by the children interacting with it.*

7

Chapter 7

# General Discussion

The ability to use hand gestures can be considered a defining property of social robots, and an important way to leverage their physical embodiment and presence. Furthermore, in human gesture studies it has been shown that a teacher's use of gestures can support students' learning process, by bringing their attention to the object of learning, by helping them understand what is communicated verbally and, in second language learning, by 'grounding' unknown words in known non-linguistic knowledge or experiences. These two premises — gestures as defining property of social robots, and the pivotal role of gestures in education — are the starting point for this thesis, and have led to the main research question: *What are the effects of robot-performed gestures in the context of second language tutoring with children, and how are these influenced by the design decisions regarding the robot's gesture production process?*

In the work that is presented in this thesis, we have addressed the research question using three different methods. Firstly, by means of a systematic review of existing literature on robot-performed gestures, we have surveyed the state of the art regarding a social robot's gesture production process, and created an overview of the effects of robot-performed gestures in various domains, including education. Secondly, we conducted three experimental studies at primary schools in the Netherlands, where children of 4–6 years old were taught English vocabulary using a social robot, and studied the effects of the robot's use of iconic gestures to support its tutoring efforts. Finally, using a semi-structured elicitation procedure in the form of a game of charades with a robot, we have collected and published a dataset of human-performed gestures, to capture natural variation that may occur when different people perform gestures for a number of concepts. This dataset can be used to inform the design of the robot's gestures, but also for gesture research in general.

The main research question was divided into eight subquestions, each of which is addressed in one or several of the chapters included in this thesis. In the following sections, we will answer these eight subquestions, discuss the relevance and implications of the results, present the limitations and avenues for future work, and answer the main research question in the general conclusion.

## 7.1 Answering the research questions

### How can we best design and implement robot-performed iconic gestures? (RQ1)

We investigated different ways to design and implement robot-performed gestures in our literature review in Chapter 2, and while setting up the studies described in Chapters 3, 4, and 6. In addition, the dataset from Chapter 5 can be used as input for designing the robot's iconic gestures. However, it is challenging to provide a conclusive answer to RQ1, as the best approach is likely to be context-dependent. For example, a manual approach, allowing more control over the robot's behavior, may be best suited for experimental research, where consistency between participants is important. However, if the robot needs to be able to engage in unconstrained, free-form dialog, a more scalable solution (e.g., automatic generation) would be desirable. Furthermore, recent developments in automatic approaches to gesture generation (e.g., synthesis) show promising results, so these approaches may see more use in the future.

In our review of existing literature (Chapter 2), we described how the design as well as the implementation, or planning, of the robot's gestures can be done either manually or automatically. Manual approaches offer more control over what the gestures will look like, and over the robot's gesturing behavior as a whole (e.g., gesture selection, frequency of gesturing). Automatic gesture design and planning (i.e., by demonstration, or gesture synthesis) on the other hand, is less labor intensive and generally results in gesturing behavior that is perceived as more natural and 'vivid' (e.g, Shimazu et al., 2018). The level of control that is offered by manual approaches provides more predictable and constrained interactions, which is why the majority of the experimental studies covered in the literature review, as well as our own studies from Chapters 3, 4, and 6 had manually designed gestures. Chapters 3 and 4 had fully scripted gesturing behavior, while in the study described in Chapter 6 we added variation by having the robot pseudorandomly perform five different gestures for the same concept.

Manually designed gestures can still be inspired by recordings of human-performed examples. This option was first explored in Chapter 4 based on an elicitation study, and then in Chapter 6 based on the dataset collected in Chapter 5. Although we did not make a direct comparison between different approaches to the design and planning of the robot's gestures, we assume that basing the design on human-performed examples, ideally from the same demographic as the one the robot will end up interacting with (in our case children), will lead to gestures that are more easily understood by the student, and subsequently should result in better learning outcomes.

Our literature review further indicated that the design of the robot's gestures can affect the way the robot is perceived, for example in terms of its human-likeness or level of enthusiasm. This is likely determined by the interplay between a robot's physical appearance and its (gesturing) behavior, although future research is needed to verify how these two factors together shape the way the robot is perceived.

A limitation of the majority of existing research on robot-performed gestures is that only few papers document their chosen design approach, and the comprehensibility of the gestures is rarely evaluated. It therefore remains an outstanding question what the best approach to designing the robot's gestures is, and this likely depends on the context in which the gestures are to be used. In our experimental studies, we did evaluate the comprehensibility of the gestures, but this was done only once. For future work, we recommend a process of iterative evaluation and refinement of the gestures, with participants from the intended target demographic, until the gestures reach their intended goal (e.g., in terms of comprehensibility, or conveying a certain personality or mood). We also urge the research field to explore ways of automating this evaluation process, or to come up with tools that can help structure human evaluations.

**What are the observed benefits of robot-performed iconic gestures in human-robot interaction, and in robot-supported education in particular, according to existing literature? (RQ2)**

Our survey of existing literature (Chapter 2) showed that robot-performed gestures can potentially serve various communicative purposes (e.g., stimulate perspective taking), influence the way the robot is perceived by others (e.g., as more human-like or likeable), increase levels of engagement with the robot, improve performance on joint tasks, and support interactees with special needs.

The literature study uncovered only limited research regarding the effects of a robot's use of (iconic) gestures in education. Studies that did focus on education showed mixed results, where math task performance did not increase (Groechel et al., 2019), and students only benefited from gestures in a learning-by-teaching task if they were already proficient at the task themselves; the gestures may have had a distracting effect on students that were not as proficient at the task (Yadollahi et al., 2018).

Several studies, one of which was in the field of education (De Carolis et al., 2019), indicate that gestures can stimulate engagement. Increased levels of engagement might lead to better learning outcomes, since students pay more attention to, and spend more time with the educational content. Engagement is also said to be indicative of a student's motivation and willingness to learn (Blumenfeld et al., 2005). One aspect that was not previously explored in the context of education is mirroring, or reenactment of the robot's gestures, although human gesture studies indicate that this may lead to a stronger contribution to learning, compared to merely observing the gestures (e.g., Tellier, 2005).

In our answer to RQ1, we stated that the design of the gestures can shape the way the robot is perceived. This can be beneficial to education, as a robot that is seen as more human-like and social, and that is liked by the student, can result in long-term engagement and relationship formation (e.g., de Graaf, 2016; van Straten et al., 2020). This, in turn, may lead children to want to keep learning with the robot for prolonged periods of time. However, it is also important to keep in mind the ethical considerations related to the use of social robots, especially when we tend to perceive these robots as human-like agents (Darling, 2017; de Graaf, 2016). In addition, basing the robots' appearance and behavior on what we know from human-human communication might prevent us from taking advantage of the 'superpowers' that robots could potentially offer (Dörrenbächer et al., 2020), for example by incorporating task-relevant sound effects (such as animal sounds when learning animal names), or by designing gestures that are different, perhaps more expressive, than those that people are physically capable of performing.

According to existing literature discussed in Chapter 2, a social robot's gestures are generally understood, although not always as well as human-performed versions. There are a number of individual differences that influence people's ability to interpret the robot's gestures: how good people are at understanding human-performed gestures, their familiarity with robots, and their age (i.e., older children are better at interpreting gestures than younger children, and adults are better than children and elderly). Integration with other modalities, such as eye gaze or facial expressions, can improve the effectiveness and clarity of the gestures. At the same time, a robot that is too lively in its social behavior can also become a distraction from the task at hand (e.g., Bourguet et al., 2020b; Kennedy et al., 2015).

**Does a robot that uses iconic gestures to support its second language tutoring efforts result in better learning outcomes than one that does not use iconic gestures? (RQ3)**

To address the need for more empirical research regarding a social robot's use of iconic gestures in education, within the L2TOR project we have conducted three studies in which a NAO robot was used to teach English vocabulary to children of 4–6 years old. In these studies, which are presented in Chapters 3, 4, and 6, we used a between-subjects design to compare between a group of children that interacted with a robot that used iconic gestures as it was training English words with them, and a group that interacted with a robot that did not use these gestures. The study described in Chapter 6 included an additional experimental condition where the robot performed five different gestures for the same concept. We found that a robot's use of iconic gestures can improve learning outcomes. However, in the study described in Chapter 3 performance only increased on a delayed test, and not on an immediate test. In addition, in the studies presented in Chapters 4 and 6, only older children appeared to benefit from the robot's gestures.

In Chapter 3, where the robot taught the names of six animals by engaging in a game of *I spy with my little eye*, we found that children in the experimental condition with iconic gestures retained more English words, as measured with a delayed post-test that was administered after at least one week, compared to children who did not receive gestures. It is possible that the lack of immediate learning gain is due to the consolidation effect, where children need time (and sleep) to process newly learned words (Axelsson et al., 2016). This is consistent with gesture studies, where the beneficial effects of gestures have been observed on a delayed, but not on an immediate post-test (e.g., McGregor et al., 2009). Consequently, it is recommended to include a delayed post-test in all studies that involve a learning task, as it might be more indicative of a child's actual learning outcomes than an immediate post-test.

Chapter 4, combined with Vogt et al. (2019), documents a long-term study of seven sessions, that included a diverse set of 34 English words, a more complex narrative-based interaction, and several different types of tasks for the child to complete together with the robot (e.g., repeating English words, enacting motion verbs). This study showed no significant effect of the robot's use of iconic gestures on students' learning outcomes. To investigate whether this was due to the complexity of the words and gestures or the tutoring interaction as a whole (e.g., regarding the large number of repetitions of the gestures in Chapter 4 compared to Chapter 3), in Chapter 6 we conducted a conceptual replication of the game of *I spy with my little eye* (Chapter 3), with a number of more abstract, complex English words and several improvements to the measurement instruments. In this study, we also found no significant effect of the use of iconic gestures on children's learning outcomes. However, on average the same number of words was learned by the children between the two studies (Chapters 3 and 6), indicating that the difficulty level of the words themselves was similar. We therefore postulate that the effectiveness of the robot's gestures might depend on the types of words that are taught (i.e., how abstract they are), and the iconicity of the matching gestures for these words. In addition, we found that age — one of the individual differences affecting the ability to interpret the robot's gestures mentioned in our answer to RQ2 — played a role, where older children in the studies from Chapters 4 and 6 benefited more from the robot's use of iconic gestures, compared to younger children. This factor is further discussed in our answer to RQ5.

**Are children more engaged with a robot that uses iconic gestures, compared to with one that does not use gestures? (RQ4)**

The effect of the robot's use of iconic gestures on engagement was studied in Chapters 3 and 6, the two studies consisting of a single session. In Chapter 3 we used an online rating study where participants were asked to provide a single rating, while in Chapter 6 engagement was annotated by the researchers using a coding scheme that distinguished between two engagement types: engagement with the (educational) task, and social engagement with the robot. Both studies showed a drop in engagement levels toward the end of the session, which is to be expected as children tend to get bored during a (repetitive) interaction. However, in both studies children showed a higher average engagement level throughout the session when the robot used iconic gestures, compared to when it did not use gestures. When distinguishing between engagement with the task and with the robot (Chapter 6), we observed that the robot's use of iconic gestures only resulted in significantly higher levels of engagement with the robot, and not with the task. This could be explained by the fact that the robot draws more attention to itself because it is moving its arms and body, resulting in more cognitive engagement. Because gestures can change the way the robot is perceived by the students (i.e., as more human-like and friendly; see our answers to RQ1 and RQ2), affective or emotional engagement could potentially be increased as well.

**What are potential factors that influence the effect of robot-performed iconic gestures on second language learning outcomes? (RQ5)**

Based on the mixed findings related to RQ3, we investigated to what extent four different factors had an influence on the effectiveness of the robot's use of iconic gestures to support its tutoring efforts. These factors were identified from gesture studies, and several of them emerged from our literature review (Chapter 2) as well. They include (1) the comprehensibility of the robot's gestures, (2) the age of the student, (3) the types of concepts that the gestures belonged to (e.g., prepositions versus motion verbs), and (4) spontaneous reenactment of the robot's gestures by the student. All four factors were investigated in Chapter 4, and age was further explored in Chapters 3 and 6. The other factors could not be studied in Chapters 3 and 6, because of the limited number of six English words included in these single-session studies, and because virtually no spontaneous reenactment took place. We found that age significantly influenced the effectiveness of the robot's use of iconic gestures, while no such effects were found for gesture comprehensibility, types of concepts, or spontaneous reenactment.

Age was found to play a significant role in two studies, the long-term study in Chapter 4 and the single session study in Chapter 6. These studies included children of approximately 5–6 years old, and 4–6 years old respectively. Older children in these studies showed better learning outcomes compared to younger children, but this only applied to the condition in which the robot used iconic gestures. It therefore seems as though younger children were unable to make use of the robot's iconic gestures to scaffold their learning process. This effect of age was not found in the first study, that is presented in Chapter 3.

We have four possible explanations for age not having an effect in the first study. First, children learned animal names, for which the gestures might have been more iconic and engaging than the concepts included in the later experiments. These gestures could have been easier for younger children to interpret. Second, we made several (albeit small) improvements to the measurement instruments before using them in the studies described in Chapters 4 and 6. In Chapter 4, we observe an effect of age in a translation task, which was not included in the studies described in Chapters 3 and 6. In Chapter 6, we improved the comprehension task from Chapter 3 by adding multiple rounds with different images (the same image from practicing with the robot, a photorealistic image, and a line drawing), and by using the L1 to introduce the questions (e.g., 'Where do you see a... [word in L2]'), instead of naming the L2 word in isolation. Third, the sample size of the study in Chapter 3 was relatively small. Finally, the inclusion of an adaptive tutoring system may have affected the difficulty level of the tasks in such a way that younger children had more cognitive effort available to make use of the robot's iconic gestures, compared to the other studies that did not include an adaptive system.

While age appears to influence learning outcomes, at least in two of the three studies, Chapter 3 shows no effect of age on overall engagement. In Chapter 6, we do see a significant effect of age on task engagement, but not on social engagement with the robot. It therefore seems that the robot's use of gestures appeals equally to children of all ages. The increase in task engagement for older children might be due to the fact that older children tend to have a longer attention span, allowing them to focus on the educational task for a longer time. However, because the effect of age on learning outcomes only applies to the experimental condition where the robot used iconic gestures, there does not appear to be a clear link between higher levels of task engagement and better learning outcomes.

The other factors — comprehensibility of the gestures, differences between types of concepts, and spontaneous reenactment — did not appear to affect children's learning gain in the long-term study (Chapter 4). However, since the study was not originally set up with the aim to investigate these factors, and children on average did not know a lot of English words at the end of the experiment, we cannot draw firm conclusions from these data alone. We did observe that children performed well on counting words, which many of them knew before the experiment, even though the gestures for these words were hard to understand. This leads us to believe that unclear gestures at least do not appear to have a detrimental effect on children's pre-existing knowledge. Furthermore, for older children that participated in the study, the robot's use of iconic gestures particularly helped them learn measurement words (e.g., *small*), and potentially also operations (e.g., *add*), although a floor effect was observed for the latter category.

An interesting aspect, that was also mentioned in our answer to RQ2, is spontaneous reenactment of the robot's gestures. This rarely happened during the single session studies with *I spy with my little eye*, while 70% of the children participating in the long-term study reenacted at least one of the robot's gestures during the first lesson. There could be a number of reasons why reenactment was more common in this study, related to the design of the robot's gestures, the physical positioning of the robot, or the implementation of the overarching tutoring interaction (e.g., having to repeat words after the robot). Literature in gesture studies (e.g., de Nooijer et al., 2013; Repetto et al., 2017) and embodied cognition (e.g., Glenberg & Gallese, 2012; Hostetter & Alibali, 2008) suggests that enactment could have a beneficial effect on students' learning outcomes, although no such effect was found in our study. In future work, we aim to uncover which design decisions can elicit reenactment of the robot's gestures by the student, and whether we can tap into the beneficial effects that are observed in the reenactment of human-performed gestures.

Additional analyses on the same data used in Chapter 4 further showed that children's selective attention and their language skills — measured by their knowledge of first language vocabulary and phonological memory — had an influence on whether they could benefit from the robot's gestures (van den Berghe et al., 2021b). Children with better selective attention performed better in the condition with iconic gestures than in the one without iconic gestures, perhaps because they had the attention and effort available to interpret the gestures, while still keeping track of the educational tasks. The study by van den Berghe et al. (2021b) also found that children with larger L1 vocabularies and better phonological memory performed better in the condition without iconic gestures than the condition with iconic gestures. This indicates that gestures can be particularly helpful for students that have relatively poor language skills, which has also been observed in human teaching scenarios (Rowe et al., 2013).

**How can we collect naturalistic human-performed examples of iconic gestures, and use these as input for designing a robot's gestures? (RQ6)** In Chapter 5, we presented a dataset of human gestures, performed by a diverse group of children and adults, collected using a game of charades with a social robot as a semi-structured elicitation procedure. Our aim with this dataset was two-fold: the recorded gestures can be used as input for a robot's gesture production and recognition processes, and for studies into human gesturing behavior (e.g., focusing on default modes of representation, or differences in gesturing between children and adults). The source code for the game of charades is made publicly available, so that the dataset can easily be extended by other researchers, for example to include new concepts, or to collect data in different (cultural) environments. Since its release, the dataset has successfully been used in gesture studies, and to inform the design of a robot's gestures.

The dataset of human-performed gestures has been used to study whether semantically similar concepts, such as *bird* and *airplane*, also share kinematic similarity (Pouw et al., 2021). This turned out to be the case, which not only provides further insight into human gesturing behavior, but it can also help robots or virtual agents produce a relevant, related gesture for a concept if no gesture was designed specifically for that concept (e.g., performing 'airplane' for the concept 'bird' if there is no gesture for ‘bird’ available). At the same time, the relation between gesture form and meaning could potentially be leveraged by gesture recognition algorithms, to detect higher level concepts (e.g., ‘means of transportation’) from gestures belonging to concrete implementations (e.g., ‘car’), on which they were not explicitly trained.

We also planned to use recorded gestures from our dataset for our final study, presented in Chapter 6, by incorporating them in a second language tutoring scenario. However, we realized that mapping the recordings directly onto the robot resulted in a substantial loss of information, such that the meaning of the gestures was difficult to infer. Therefore, we manually recreated the gestures based on examples from the dataset (as suggested in our answer to RQ1). These gestures were evaluated using an online questionnaire, and gestures that scored low on comprehensibility were revised. By basing the gestures on examples from the dataset, a number of which were performed by children, we believe that they matched better with the preferred gesture forms of children participating in our study, compared to designing the gestures from the researchers’ frame of reference. However, we have not yet made a direct comparison between these different approaches to the design of the robot’s gestures. It is also worth noting that the dataset that was used in this case consisted of pantomime (silent) gestures, while researchers such as McNeill, 1992 have argued that gesture and speech should be understood (and therefore, perhaps, also recorded) together as an integrated system.

**Do gestures contribute more to learning performance when multiple gestures are used for the same concept, highlighting different salient features of this concept, compared to a single gesture for each concept? (RQ7)**

Research in gesture studies (e.g., Ortega & Özyürek, 2016), and the recordings in our dataset from Chapter 5 show that there is variation in how people produce gestures for a particular concept. This might be caused by differences in mental representations of objects and concepts (cf. Piaget & Cook, 1952). We therefore postulate that, in education, variation in the cues that support the learning process, focusing on different aspects of the object of learning, might lead to increased learning outcomes. This finds further support in variation theory (Marton & Booth, 2013). In Chapter 6, we therefore set out to investigate whether varation in the robot’s gesturing behavior has an effect on children’s learning outcomes. This was done by means of a study with three experimental conditions: no gestures, repeated gestures (one for each concept), or varied gestures (five for each concept). Children again played the game of *I spy with my little eye* with the robot, in a single session.

The results showed no significant difference in learning outcomes between the conditions. However, the same effect of age that was presented in our answer to RQ5 for the repeated gestures condition was observed for the condition with varied gestures: Older children in the study on average learned more words than younger children, but only for the two conditions with repeated and varied iconic gestures.

**Does variation in the robot's gesture repertoire result in higher levels of engagement with the robot or the task, compared to a single gesture for each concept? (RQ8)**

Based on existing literature (e.g., Tanaka et al., 2007), we assumed that children would feel more engaged with a robot that shows more variation in its behavior. In Chapter 6, we found that varied gestures had the same effect as repeated gestures on engagement, where the average level of engagement with the robot was higher for both conditions with gestures, compared to the condition without gestures, but no significant difference was found between repeated and varied gestures. There was no effect of varied or repeated gestures on task engagement.

To summarize the answers to RQ7 and RQ8, the results of the study presented in Chapter 6 do not point toward any benefits of the robot's use of varied gestures in the context of education, compared to repeating the same gesture. However, there are also no apparent drawbacks to including a number of different gestures. It might be the case that a robot that varies its behavior is perceived more positively than one that shows repetitive behavior, particularly if the student is to engage in multiple sessions with the robot. Future long-term studies are needed to further investigate these effects. Because there is little research focusing on the role of various forms of variation in educational settings, it might be worthwhile to take a step back and investigate this topic with human-performed gestures, before studying this in the context of human-robot interaction. In addition, studies could involve older children or adults, as they are more capable of reflecting upon and verbalizing their experiences, which will provide qualitative data on how variation is perceived by people interacting with the robot.

## 7.2 Implications and recommendations

Our contribution to research into the role of robot-performed gestures in education is three-fold. We have provided (1) a comprehensive overview of the state of the art in robot-performed gestures, (2) empirical findings regarding the effects of robot-performed iconic gestures in the context of second language tutoring with children, and (3) a dataset of human-performed gestures.

In our review of existing literature in the broader domain of robot-performed gestures (beyond iconic gestures for education), presented in Chapter 2, we have provided an overview of the state of the art in the field, and concluded with a list of ten outstanding questions and four methodological suggestions. These form our concrete recommendations for studying the role of robot-performed gestures in human-robot interactions, and can serve as guidelines for future research.

Chapters 3, 4, and 6 present the first studies into the effects of robot-performed iconic gestures in the context of second language tutoring. In doing so, we have elaborately described the process of designing the tutoring interactions and the robot's gestures, and the systems and gestures that were developed have been made publicly available to support future work in this emerging field of research.

Our findings indicate that iconic gestures can be considered a way for robots to make use of their physical presence, and as part of their socially intelligent behavior. In second language learning, we found that the robot's use of iconic gestures can help with communicating the educational content, resulting in better learning outcomes, as well as facilitating interest, which led to higher levels of engagement with the robot. This can be considered an indication of a greater likelihood of, and tendency toward building a lasting relationship with a robot that uses gestures.

The main factor influencing the success of the robot's iconic gestures appears to be the students' age: we found that older children in our study (of approximately 5.5–6 years old) were able to learn more with help from the robot's iconic gestures, while younger children did not seem to benefit from them. This knowledge can be used to inform the design of robot tutoring interactions, and particularly the robot's iconic gestures, in the future.

Variation in the robot's gestures has previously been underresearched, and was first explored in Chapter 6 of this thesis. Although we did not observe any benefits of varied gestures compared to repeated gestures, this remains an interesting topic for future studies from the perspectives of education, gesture studies, and social robotics alike. Robots can prove to be useful tools in these future studies, because their behavior can be made consistent across participants, and studies can easily be replicated. This is also why we used a robot confederate in the gesture elicitation study (Chapter 5).

Finally, in Chapter 5 we introduced a dataset of recorded gestures, performed by a large and diverse group of participants in our gameful elicitation study, which has been made publicly available. It can easily be extended to include more concepts or more recordings. The gesture recordings are relevant for gesture research, and they can be used as input for the gesture production and recognition processes of virtual agents and robots.

## 7.3 Limitations and future work

To ensure that the studies conducted within the context of this thesis and the L2TOR project remained feasible, and to minimize the influence of confounding variables, several concessions had to be made. At the same time, these provide avenues for future work. We have identified five limitations, which we will briefly discuss below.

### Use of the NAO robot

All of the studies discussed in this thesis used the same social robot, the SoftBank Robotics NAO V5. We chose to use this robot because of its (relative) availability and affordability, and because we expected that its appearance would appeal to younger children. Its popularity in research on social robots ensures that our findings can be positioned within the broader research field. At the same time, the use of only one type of robot can be seen as a limitation, as the literature review (Chapter 2) indicated that the robot's appearance could also influence the effectiveness of the robot's gestures. In addition, the gestures in our studies were designed specifically for the NAO robot. Other robots may have different motor degrees of freedom, and different features (e.g., five fingers instead of three), which will have an impact on their gesturing behavior. In future work, we therefore recommend to compare between different robot platforms, to verify whether our results generalize to other robots. This is also proposed as a methodological suggestion at the end of Chapter 2, to ensure external validity of studies on robot-performed gestures.

### Technical limitations

While developing the intelligent tutoring system, we encountered a number of technical limitations, particularly related to the robot's sensing capabilities (e.g., detecting physical objects, automatic speech recognition). To work around these limitations, we decided to introduce a tablet device on which the educational content was presented, and limited the amount of verbal interaction with the robot. The addition of a tablet device may have increased cognitive load for the students, as they had to split their attention between the tablet and the robot. This, in turn, may have made it more challenging for children to make use of the robot's gestures. The fact that children with better selective attention performed better in the experimental condition with iconic gestures compared to the condition without iconic gestures further supports this point (van den Berghe et al., 2021b). Children might have more cognitive effort to spend on observing the robot and its gestures if simple physical objects, or no objects at all, would have been used. In the future, we therefore intend to study interactions that involve only the child and the robot.

### Outstanding questions regarding robot-performed gestures

A further limitation is that we were not able to address all of the ten outstanding questions, presented in our literature review (Chapter 2), in the current thesis. This is because some of the outstanding questions rely on future developments (i.e., in sensor technology and AI), because of the aforementioned limitation of only using the NAO robot, or because we had to limit the scope of our research, to keep the focus on second language learning. However, we did contribute to six of the outstanding questions (1, 2, 4, 5, 7, and 8), as discussed at the end of Chapter 2. The source code and materials of our studies can be used to continue our line of research. We look forward to discussing these outstanding questions and our contributions with other researchers in the field.

### Research with young children

Our research focused on second language tutoring for relatively young children, of 4–6 years old — the age group attending the first two years of primary school in the Netherlands. This age was chosen because children's academic success is said to depend on early instruction of language skills (Hoff, 2013; Nikolov & Djigunović, 2006), and because mastering a foreign language is considered a useful skill for their future (European Commission, 2012).

There are a number of challenges when conducting research with young children. For example, it can be difficult for them to reflect upon and verbalize their experiences in a detailed, qualitative manner (Markopoulos et al., 2008). As a result, we had to rely on adults to, for example, rate the clarity of the gestures or the engagement levels of the children. We also observed large individual differences in our studies with children, which may limit the generalizability of the findings, particularly in our first study (Chapter 3) that had a relatively small sample. In addition, children in general did not learn many new words in our studies, making it more difficult to study the influence of various factors, such as the robot's use of gestures, on these learning outcomes. The effect of age, that was observed within this age group, may no longer be a factor if we were to focus on older children or adults. In future work we aim to broaden the age group of our participants, to get additional qualitative and fine-grained insights regarding the use of social robots as second language tutors.

**Focus on second language vocabulary**

Existing research in human gesture studies indicates that iconic gestures can be beneficial to education, particularly in the domain of second language learning (e.g., Repetto et al., 2017; Tellier, 2008). This is potentially due to the 'grounding' effect, where gestures can be used to link new linguistic concepts to familiar non-linguistic knowledge and experiences (Barsalou, 2008). Because we focused solely on second language learning, it is possible that our findings do not generalize to other educational domains. However, research with social robots in related fields does show promising results (e.g., Bremner et al., 2011; Huang & Mutlu, 2013; van Dijk et al., 2013).

In addition, our focus with this work was on vocabulary training, but there are several other aspects of language learning where robots could offer support, such as practicing with having conversations in a second language. People might be less anxious when talking to a robot in a second language, compared to talking with another person (Alemi et al., 2015). Therefore, we postulate that robots could have additional benefits for aspects of language learning other than vocabulary training. However, because we designed the interaction for young children that had little to no pre-existing vocabulary knowledge, and to ensure that the interaction would be the same for all children, we decided to limit the content of the tutoring system to short vocabulary terms for the present studies.

## 7.4 The future of robots in education

The goal of the L2TOR project, and by extension of the work that was conducted in the context of this thesis, was to create an intelligent tutoring system to help children learn a second language together with a social robot. Ideally, by the end of the project there would be a system that was ready to be handed over to schools, to be used in practice. While the first part of this plan worked — children successfully learned

English words by engaging in the tutoring interaction — the second part turned out to be more challenging than expected. Because of various technical limitations, and because a certain amount of technical knowledge is still required to program the robot and to generate new content, we were unable to deliver a plug-and-play solution that was ready to be handed over to schools, that could easily be extended with new content by teachers, and that could take full advantage of the potential that social robots have to offer.

This experience within our project appears to echo the general sentiment about social robotics, where it is said that we have entered a *social robotics winter* (Henschel et al., 2020). This refers to the disillusionment that follows a period of (over)inflated expectations. The Gartner hype cycle (Fenn & Raskino, 2008) describes a similar development over time for various emerging technologies, such as social robots. According to the hype cycle, all technologies go through a process of having inflated expectations, after which a phase of enlightenment brings us to a plateau of productivity, at which point we understand what a technology can and cannot do for us. According to Gartner, the expectations surrounding 'smart robots' were still estimated to be on the rise in 2020, with 5–10 years to go before the plateau of productivity would be reached[1].

Since all new technologies appear to follow the same pattern, although at different paces, we believe there is a lot we can learn from other tools that have previously been used to innovate education, such as tablet devices. This includes, for example, making robots more affordable and accessible to a wider audience. This audience should include teachers, educational publishers, parents, and the students themselves. They should be given the tools to create additional content for their robots, so that as a society we can together explore the role that robots can and should take up in our lives. It is likely that this role will be similar to that of a tablet: as a tool to support teaching, but certainly not replace teachers, with the added benefit of having a social and physical presence in the context of the student.

My personal vision of the ideal social robot in education is one that is affordable, open source, and tailored to the individual. Similar to the NAO robot, it has the appearance of a toy with human-like features. However, it has a smaller form factor, so that it can sit on the student's desk and be carried around in a backpack. The robot can support the student in all educational activities, at school and at home. By

[1] https://web.archive.org/web/20210813185706/https://www.gartner.com/smarterwithgartner/2-megatrends-dominate-the-gartner-hype-cycle-for-artificial-intelligence-2020/

tracking the student's progress across different subjects, it is able to provide guidance on a metacognitive level (e.g., regarding learning strategies), it can keep track of the student's schedule, and indicate when it is time to take a break from studying. In other words, it becomes a personal companion to the individual student. There are two concerns, however, that I would urge to explore before making this vision a reality. First, there are ethical implications to consider. For example, how does giving a robot to children affect their (social) development? How would children respond if their companion robot, that they may have built a relationship with, breaks down or has a bug? Second, it might be difficult for a small form factor robot to perform gestures that are elaborate enough to provide the benefits that we observed with the NAO robot in our studies, or to be able to use sign language to communicate.

## 7.5 General conclusion

The aim of this thesis was to investigate the effects of a social robot's use of iconic gestures to support its second language tutoring efforts with children of 4–6 years old. We addressed the following research question: *What are the effects of robot-performed gestures in the context of second language tutoring with children, and how are these influenced by the design decisions regarding the robot's gesture production process?*

Our studies showed that a robot's use of iconic gestures while teaching children a second language can result in better learning outcomes. For gestures that were less iconic, this only applied to older children in our samples (of approximately 5.5–6 years old). Children also tend to be more engaged with a robot that uses gestures, compared to one that does not gesture. Throughout the different studies we have reflected upon, and made improvements to, the process of designing and evaluating the robot's gestures.

In our review of existing literature, we found that design decisions regarding the robot's gesture production process, such as making sure the gestures are congruent with what is communicated verbally, do appear to influence the effectiveness of the robot's gestures. In our studies, we have based the gestures on human-performed examples, included variation in the robot's gesturing behavior, and observed spontaneous reenactment of the robot's gestures. None of these factors appeared to have an effect on children's learning outcomes, although future research is needed to be able to draw firm conclusions. Our dataset of collected human-performed gestures can be used to identify other factors, such as differences between adults and children, that may inform the design of robot-performed gestures in the future.

At the start of our project, expectations of what social robots could do — in education and otherwise — ran high. Over the last few years, we have sometimes observed signs of disillusionment that may imply that we are entering a social robotics winter, where the capabilities of robots that people currently interact with do not meet the expectations that were set out when social robots first became available, and that are implied by their human-like appearance. However, we believe that social robots have a lot to offer in education, especially if we follow the path of accessibility, availability, and openness: making sure that everyone is able to own, and create content for the social robots of the future. While the hype surrounding social robots may be winding down, a more realistic image of what these robots can and cannot do for us will take shape. With the studies presented in this thesis, we aim to provide a realistic account of current developments in social robotics for education, and the role that gestures could play in supporting educational tasks, to help complete this image.

We expect that the process of introducing robots to schools and to our lives may follow a similar trajectory to that of other technologies that preceded them, such as computers and tablets. However, compared to other tools that are currently used in education, social robots can have the added benefit of being present, as social beings, in our physical environment. We have observed that their ability to use gestures can improve children's learning outcomes, although more research is needed to explore the design space of robot-performed gestures, in order to optimally make use of their contribution to a social robot's tutoring efforts. With this vision for the future of (gesturing) robots in education, we believe that we can leave the social robotics winter behind us, and enter a blossoming social robotics spring.

# Bibliography

Abdul Jalil, S. B., Huang, J., Markovich, M., Osburn, B., Barley, M., & Amor, R. (2012). Avatars at a meeting. *ACM International Conference Proceeding Series*, 84–87. https://doi.org/10.1145/2379256.2379270

Abramov, O., Kern, F., Koutalidis, S., Mertens, U., Rohlfing, K., & Kopp, S. (2021). The relation between cognitive abilities and the distribution of semantic features across speech and gesture in 4-year-olds. *Cognitive Science*, *45*(7), e13012. https://doi.org/10.1111/cogs.13012

Admoni, H., & Scassellati, B. (2017). Social eye gaze in human-robot interaction: A review. *J. Hum.-Robot Interact.*, *6*(1), 25–63. https://doi.org/10.5898/JHRI.6.1.Admoni

Admoni, H., Weng, T., Hayes, B., & Scassellati, B. (2016). Robot nonverbal behavior improves task performance in difficult collaborations. *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*, 51–58. https://doi.org/10.1109/HRI.2016.7451733

Ahmad, M. I., Mubin, O., & Orlando, J. (2016a). Children views' on social robot's adaptations in education. *Proceedings of the 28th Australian Computer-Human Interaction Conference, OzCHI 2016*, 145–149. https://doi.org/10.1145/3010915.3010977

Ahmad, M. I., Mubin, O., & Orlando, J. (2016b). Understanding behaviours and roles for social and adaptive robots in education: Teacher's perspective. *HAI 2016 - Proceedings of the 4th International Conference on Human Agent Interaction*, 297–304. https://doi.org/10.1145/2974804.2974829

Ahn, H. S., Lee, D. W., & MacDonald, B. (2013). Development of a human-like narrator robot system in EXPO. *IEEE Conference on Robotics, Automation and Mechatronics, RAM - Proceedings*, 7–12. https://doi.org/10.1109/RAM.2013.6758551

Aizawa, M., & Umemuro, H. (2021). Behavioral design of guiding agents to encourage their use by visitors in public spaces. *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 247–251. https://doi.org/10.1145/3434074.3447169

Akalin, N., Uluer, P., Kose, H., & Ince, G. (2013). Humanoid robots communication with participants using sign language: An interaction based sign language game. *Proceedings of IEEE Workshop on Advanced Robotics and its Social Impacts, ARSO*, 181–186. https://doi.org/10.1109/ARSO.2013.6705526

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*(4), 523–535.

Ali, W., & Williams, A. B. (2020). Evaluating the effectiveness of nonverbal communication in human-robot interaction. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 99–100. https://doi.org/10.1145/3371382.3378354

Alibali, M. W., & Nathan, M. J. (2007). Teachers' gestures as a means of scaffolding students' understanding: Evidence from an early algebra lesson. *Video Research in the Learning Sciences*, *39*(5), 349–366. https://doi.org/10.1111/j.1467-8535.2008.00890_7.x

Aloba, A., Flores, G., Woodward, J., Shaw, A., Castonguay, A., Cuba, I., Dong, Y., Jain, E., & Anthony, L. (2018). Kinder-Gator: The UF Kinect database of child and adult motion. *Eurographics (Short Papers)*, 13–16.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175–185.

Aly, A., & Tapus, A. (2012). Prosody-driven robot arm gestures generation in human-robot interaction. *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 257–258. https://doi.org/10.1145/2157689.2157783

Aly, A., & Tapus, A. (2013). A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction. *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, 325–332.

Aly, A., & Tapus, A. (2015). Multimodal adapted robot behavior synthesis within a narrative human-robot interaction. *IEEE International Conference on Intelligent Robots and Systems*, *2015-Decem*, 2986–2993. https://doi.org/10.1109/IROS.2015.7353789

Aly, A., & Tapus, A. (2016). Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human–robot interaction. *Autonomous Robots*, *40*(2), 193–209. https://doi.org/10.1007/s10514-015-9444-1

Aly, A., & Tapus, A. (2020). On designing expressive robot behavior: The effect of affective cues on interaction. *SN Computer Science*, *1*(6), 1–17. https://doi.org/10.1007/s42979-020-00263-3

Anzalone, S. M., Boucenna, S., Ivaldi, S., & Chetouani, M. (2015). Evaluating the engagement with social robots. *International Journal of Social Robotics*, *7*(4), 465–478.

Arachchige, K. G. K., Loureiro, I. S., Blekic, W., Rossignol, M., & Lefebvre, L. (2021). The role of iconic gestures in speech comprehension: An overview of various methodologies. *Frontiers in Psychology*, *12*.

Argall, B. D., Chernova, S., Veloso, M., & Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, *57*(5), 469–483.

Arici, T., Celebi, S., Aydin, A. S., & Temiz, T. T. (2014). Robust gesture recognition using feature pre-processing and weighted dynamic time warping. *Multimedia Tools and Applications*, *72*(3), 3045–3062.

Asselborn, T., Johal, W., & Dillenbourg, P. (2017). Keep on moving! exploring anthropomorphic effects of motion during idle moments. *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 897–902.

Augello, A., & Pilato, G. (2019). An annotated corpus of stories and gestures for a robotic storyteller. *Proceedings - 3rd IEEE International Conference on Robotic Computing, IRC 2019*, 630–635. https://doi.org/10.1109/IRC.2019.00127

Augustine, A. C., Ryusuke, M., Liu, C., Ishi, C. T., & Ishiguro, H. (2020). Generation and evaluation of audio-visual anger emotional expression for android robot. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 96–98. https://doi.org/10.1145/3371382.3378282

Aussems, S., & Kita, S. (2019). Seeing iconic gestures while encoding events facilitates children's memory of these events. *Child Development*, *90*(4), 1123–1137. https://doi.org/10.1111/cdev.12988

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*, 1192.

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, *3*(1), 41–52.

Bao, Y., & Cuijpers, R. H. (2017). On the imitation of goal directed movements of a humanoid robot. *International Journal of Social Robotics*, *9*(5), 691–703. https://doi.org/10.1007/s12369-017-0417-8

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, *27*(3), 387–414.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*, 617–645.

Bartneck, C., & Forlizzi, J. (2004). A design-centred framework for social human-robot interaction. *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)*, 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.

Beattie, G. (2003). *Visible thought: The new psychology of body language*. Psychology Press.

Belpaeme, T., Kennedy, J., Baxter, P., Vogt, P., Krahmer, E. J., Kopp, S., Bergmann, K., Leseman, P., Küntay, A. C., Göksun, T., et al. (2015). L2TOR-second language tutoring using social robots. *Proceedings of the ICSR 2015 WONDER Workshop*.

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

Bennewitz, M., Faber, F., Joho, D., Schreiber, M., & Behnke, S. (2005). Towards a humanoid museum guide robot that interacts with multiple persons. *Proceedings of 2005 5th IEEE-RAS International Conference on Humanoid Robots, 2005*, 418–423. https://doi.org/10.1109/ICHR.2005.1573603

Bergmann, K., & Macedonia, M. (2013). A virtual agent as vocabulary trainer: Iconic gestures help to improve learners' memory performance. *International Workshop on Intelligent Virtual Agents*, 139–148.

Bethel, C. L., & Murphy, R. R. (2010). Review of human studies methods in HRI and recommendations. *International Journal of Social Robotics*, *2*(4), 347–359.

Blatchford, P., & Russell, A. (2020). *Rethinking class size: The complex story of impact on teaching and learning*. UCL Press.

Blumenfeld, P. C., Kempler, T. M., & Krajcik, J. S. (2005). Motivation and cognitive engagement in learning environments. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 475–488). Cambridge University Press. https://doi.org/10.1017/CBO9780511816833.029

Booth, A. E., McGregor, K. K., & Rohlfing, K. J. (2008). Socio-pragmatics and attention: Contributions to gesturally guided word learning in toddlers. *Language Learning and Development*, *4*(3), 179–202. https://doi.org/10.1080/15475440802143091

Bosker, H. R., & Peeters, D. (2021). Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1943), 20202419. https://doi.org/10.1098/rspb.2020.2419

Bourguet, M.-L., Jin, Y., Shi, Y., Chen, Y., Rincon-Ardila, L., & Venture, G. (2020a). Social robots that can sense and improve student engagement. *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 127–134. https://doi.org/10.1109/TALE48869.2020.9368438

Bourguet, M.-L., Xu, M., Zhang, S., Urakami, J., & Venture, G. (2020b). The impact of a social robot public speaker on audience attention. *Proceedings of the 8th International Conference on Human-Agent Interaction*, 60–68. https://doi.org/10.1145/3406499.3415073

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Bragdon, A., Uguray, A., Wigdor, D., Anagnostopoulos, S., Zeleznik, R., & Feman, R. (2010). Gesture play: Motivating online gesture learning with fun, positive reinforcement and physical metaphors. *ACM International Conference on Interactive Tabletops and Surfaces*, 39–48.

Bragdon, A., Zeleznik, R., Williamson, B., Miller, T., & LaViola Jr, J. J. (2009). Gesture-Bar: Improving the approachability of gesture-based interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2269–2278.

Brand, R. J., & Shallcross, W. L. (2008). Infants prefer motionese to adult-directed action. *Developmental Science*, *11*(6), 853–861. https://doi.org/10.1111/j.1467-7687.2008.00734.x

Breazeal, C. (2004). *Designing sociable robots*. MIT press.

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. *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 708–713.

Bremner, P., Celiktutan, O., & Gunes, H. (2016). Personality perception of robot avatar tele-operators. *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*, 141–148. https://doi.org/10.1109/HRI.2016.7451745

Bremner, P., & Leonards, U. (2015a). Efficiency of speech and iconic gesture integration for robotic and human communicators-a direct comparison. *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 1999–2006.

Bremner, P., & Leonards, U. (2015b). Speech and gesture emphasis effects for robotic and human communicators: A direct comparison. *ACM/IEEE International Conference on Human-Robot Interaction*, *2015-March*, 255–262. https://doi.org/10.1145/2696454.2696496

Bremner, P., & Leonards, U. (2016). Iconic gestures for robot avatars, recognition and integration with speech. *Frontiers in Psychology*, *7*, 183.

Bremner, P., Pipe, A. G., Fraser, M., Subramanian, S., & Melhuish, C. (2009). Beat gesture generation rules for human-robot interaction. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 1029–1034. https://doi.org/10.1109/ROMAN.2009.5326136

Bremner, P., Pipe, A. G., Melhuish, C., Fraser, M., & Subramanian, S. (2011). The effects of robot-performed co-verbal gesture on listener behaviour. *IEEE-RAS International Conference on Humanoid Robots*, 458–465. https://doi.org/10.1109/Humanoids.2011.6100810

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of standardized stimuli (BOSS) phase II: 930 new normative photos. *PLoS One*, *9*(9), e106953.

Burns, R., Jeon, M., & Park, C. H. (2018). Robotic motion learning framework to promote social engagement. *Applied Sciences (Switzerland)*, *8*(2). https://doi.org/10.3390/app8020241

Cabibihan, J. J., So, W. C., & Pramanik, S. (2012). Human-recognizable robotic gestures. *IEEE Transactions on Autonomous Mental Development*, *4*(4), 305–314. https://doi.org/10.1109/TAMD.2012.2208962

Cabrera, M. E., Novak, K., Foti, D., Voyles, R., & Wachs, J. P. (2017). What makes a gesture a gesture? neural signatures involved in gesture recognition. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 748–753.

Cabrera, M. E., & Wachs, J. P. (2017). A human-centered approach to one-shot gesture learning. *Frontiers in Robotics and AI*, *4*, 8.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7291–7299.

Carter, E. J., Mistry, M. N., Carr, G. P. K., Kelly, B. A., & Hodgins, J. K. (2014). Playing catch with robots: Incorporating social gestures into physical interactions. *IEEE RO-MAN 2014 - 23rd IEEE International Symposium on Robot and Human Interactive Communication: Human-Robot Co-Existence: Adaptive Interfaces* *and Systems for Daily Life, Therapy, Assistance and Socially Engaging Interactions*, 231–236. https://doi.org/10.1109/ROMAN.2014.6926258

Cassell, J., Vilhjálmsson, H. H., & Bickmore, T. (2004). BEAT: The behavior expression animation toolkit. *Life-like characters* (pp. 163–185). Springer.

Chang, C.-W., Lee, J.-H., Chao, P.-Y., Wang, C.-Y., & Chen, G.-D. (2010). Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school. *Journal of Educational Technology & Society*, *13*(2), 13–24.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893.

Chiat, S. (2015). Nonword repetition. *Methods for assessing multilingual children: Disentangling bilingualism from language impairment*, 125–150.

Chidambaram, V., Chiang, Y. H., & Mutlu, B. (2012). Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues. *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 293–300. https://doi.org/10.1145/2157689.2157798

Christoforakos, L., Gallucci, A., Surmava-Große, T., Ullrich, D., & Diefenbach, S. (2021). Can robots earn our trust the same way humans do? a systematic exploration of competence, warmth and anthropomorphism as determinants of trust development in HRI. *Frontiers in Robotics and AI*, *8*, 79.

Cifuentes, C. A., Pinto, M. J., Céspedes, N., & Múnera, M. (2020). Social robots in therapy and care. *Current Robotics Reports*, 1–16.

Claret, J. A., Venture, G., & Basañez, L. (2017). Exploiting the robot kinematic redundancy for emotion conveyance to humans as a lower priority task. *International Journal of Social Robotics*, *9*(2), 277–292. https://doi.org/10.1007/s12369-016-0387-2

Clark, H. H. (1996). *Using language*. Cambridge university press.

Connell, S., Kuo, P.-Y., Liu, L., & Piper, A. M. (2013). A wizard-of-oz elicitation study examining child-defined gestures with a whole-body interface. *Proceedings of the 12th International Conference on Interaction Design and Children*, 277–280.

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition*, *106*(2), 1047–1058. https://doi.org/10.1016/j.cognition.2007.04.010

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278.

Craenen, B., Deshmukh, A., Foster, M. E., & Vinciarelli, A. (2018). Shaping gestures to shape personalities: The relationship between gesture parameters, attributed personality traits and godspeed scores. *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 699–704.

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with autotutor. *Journal of Educational Media*, *29*(3), 241–250.

Cravotta, A., Busà, M. G., & Prieto, P. (2019). Effects of encouraging the use of gestures on speech. *Journal of Speech, Language, and Hearing Research*, *62*(9), 3204–3219.

Dael, N., Goudbeek, M., & Scherer, K. R. (2013). Perceived gesture dynamics in nonverbal expression of emotion. *Perception*, *42*(6), 642–657.

Dargue, N., & Sweller, N. (2018). Not all gestures are created equal: The effects of typical and atypical iconic gestures on narrative comprehension. *Journal of Nonverbal Behavior*, *42*(3), 327–345.

Darling, K. (2017). 'Who's Johnny?' Anthropomorphic framing in human-robot interaction, integration, and policy. In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence*. Oxford University Press.

De Carolis, B., Palestra, G., Della Penna, C., Cianciotta, M., & Cervelione, A. (2019). Social robots supporting the inclusion of unaccompanied migrant children: Teaching the meaning of culture-related gestures. *Journal of E-Learning and Knowledge Society*, *15*(2), 43–57. https://doi.org/10.20368/1971-8829/1636

de Graaf, M. M. (2016). An ethical evaluation of human–robot relationships. *International Journal of Social Robotics*, *8*(4), 589–598. https://doi.org/10.1007/s12369-016-0368-5

de Graaf, M. M., & Ben Allouch, S. (2013). Exploring influencing variables for the acceptance of social robots. *Robotics and Autonomous Systems*, *61*(12), 1476–1486. https://doi.org/10.1016/j.robot.2013.07.007

de Haas, M., Vogt, P., & Krahmer, E. (2020). The effects of feedback on children's engagement and learning outcomes in robot-assisted second language learning. *Frontiers in Robotics and AI*. https://doi.org/10.3389/frobt.2020.00101

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. https://doi.org/10.1016/j.actpsy.2013.05.013

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. *Workshop on the Challenges of Working on Social Robots that Collaborate with People, at the ACM CHI Conference on Human Factors in Computing Systems (CHI2019).*

de Wit, J., Schodde, T., Willemsen, B., Bergmann, K., de Haas, M., Kopp, S., Krahmer, E., & Vogt, P. (2017). Exploring the effect of gestures and adaptive tutoring on children's comprehension of L2 vocabularies. *Proceedings of the Workshop R4L at ACM/IEEE HRI 2017.*

DeLoache, J. S. (2004). Becoming symbol-minded. *Trends in Cognitive Sciences*, *8*(2), 66–70. https://doi.org/10.1016/j.tics.2003.12.004

Dennett, D. C. (1987). *The intentional stance*. MIT press.

DePalma, N., Smith, J., Chernova, S., & Hodgins, J. (2021). Toward a one-interaction data-driven guide: Putting co-speech gesture evidence to work for ambiguous route instructions. *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 505–509. https://doi.org/10.1145/3434074.3447223

Deshmukh, A., Craenen, B., Vinciarelli, A., & Foster, M. E. (2018). Shaping robot gestures to shape users' perception: The effect of amplitude and speed on Godspeed ratings. *HAI 2018 - Proceedings of the 6th International Conference on Human-Agent Interaction*, 293–300. https://doi.org/10.1145/3284432.3284445

Dijkstra, K., & Post, L. (2015). Mechanisms of embodiment. *6*(OCT), 1525. https://doi.org/10.3389/fpsyg.2015.01525

Dörrenbächer, J., Löffler, D., & Hassenzahl, M. (2020). Becoming a robot-overcoming anthropomorphism with techno-mimesis. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Drijvers, L., & Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research*, *60*(1), 212–222.

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

Duffy, B. R., & Joue, G. (2000). Intelligent robots: The question of embodiment. *Proc. of the Brain-Machine Workshop.*

Eisenbeiss, S. (2010). Production methods in language acquisition research. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 11–34). John Benjamins Publishing Company Amsterdam.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica 1*, 49–98.

Émond, C., Lewis, L., Chalghoumi, H., & Mignerat, M. (2020). A comparison of NAO and Jibo in child-robot interaction. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 192–194. https://doi.org/10.1145/3371382.3378234

European Commission. (2012). Special eurobarometer 386: Europeans and their languages. *URL: https://web.archive.org/web/20120710072925/http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf*.

Fenn, J., & Raskino, M. (2008). *Mastering the hype cycle: How to choose the right innovation at the right time.* Harvard Business Press.

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

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*(3-4), 143–166.

Fridin, M. (2014). Kindergarten social assistive robot: First meeting and ethical issues. *Computers in Human Behavior*, *30*, 262–272.

Ghosh, B., Dhall, A., & Singla, E. (2019). Automatic speech-gesture mapping and engagement evaluation in human robot interaction. *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1–7.

Gielniak, M. J., Liu, C. K., & Thomaz, A. L. (2011). Task-aware variations in robot motion. *Proceedings - IEEE International Conference on Robotics and Automation*, 3921–3927. https://doi.org/10.1109/ICRA.2011.5980348

Gielniak, M. J., & Thomaz, A. L. (2012). Enhancing interaction through exaggerated motion synthesis. *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 375–382.

Glenberg, A. M., & Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex*, *48*(7), 905–922.

Goldin-Meadow, S. (2000). Beyond words: The importance of gesture to researchers and learners. *Child Development*, *71*(1), 231–239. https://doi.org/10.1111/1467-8624.00138

Goldin-Meadow, S. (2005). *Hearing gesture: How our hands help us think*. Harvard University Press.

González, F. J., Perez-Uribe, A., Satizábal, H. F., & López, J. A. (2019). DCGAN model used to generate body gestures on a human-humanoid interaction system. *Communications in Computer and Information Science*, *1096 CCIS*, 103–115. https://doi.org/10.1007/978-3-030-36211-9_9

Gordon, G., & Breazeal, C. (2015). Bayesian active learning-based robot tutor for children's word-reading skills. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 1343–1349.

Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., Das, M., & Breazeal, C. (2016). Affective personalization of a social robot tutor for children's second language skills. *Thirtieth AAAI Conference on Artificial Intelligence*.

Goto, M., Yokoyama, M., & Matsuura, Y. (2020). Impression evaluation of presentation by a communication robot in an actual exhibition. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 218–220. https://doi.org/10.1145/3371382.3378264

Groechel, T., Shi, Z., Pakkar, R., & Mataric, M. J. (2019). Using socially expressive mixed reality arms for enhancing low-expressivity robots. *2019 28th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2019*. https://doi.org/10.1109/RO-MAN46459.2019.8956458

Gulzar, K., & Kyrki, V. (2015). See what I mean — probabilistic optimization of robot pointing gestures. *IEEE-RAS International Conference on Humanoid Robots*, *2015-Decem*, 953–958. https://doi.org/10.1109/HUMANOIDS.2015.7363484

Hald, L. A., de Nooijer, J., van Gog, T., & Bekkering, H. (2016). Optimizing word learning via links to perceptual and motoric experience. *Educational Psychology Review*, *28*(3), 495–522. https://doi.org/10.1007/s10648-015-9334-2

Ham, J., Cuijpers, R. H., & Cabibihan, J. J. (2015). Combining robotic persuasive strategies: The persuasive power of a storytelling robot that uses gazing and gestures. *International Journal of Social Robotics*, *7*(4), 479–487. https://doi.org/10.1007/s12369-015-0280-4

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*(4), 159–168.

Hanfstingl, B., Benke, G., & Zhang, Y. (2019). Comparing variation theory with Piaget's theory of cognitive development: More similarities than differences? *Educational Action Research*, 1–16.

Hasegawa, D., Cassell, J., & Araki, K. (2010). The role of embodiment and perspective in direction-giving systems. *2010 AAAI Fall Symposium Series.*

Hato, Y., Satake, S., Kanda, T., Imai, M., & Hagita, N. (2010). Pointing to space: Modeling of deictic interaction referring to regions. *5th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2010*, 301–308. https://doi.org/10.1145/1734454.1734559

Haviland, J. B. (2000). Pointing, gesture spaces, and mental maps. *Language and gesture*, *2*, 13.

Hayes, C. J., Crowell, C. R., & Riek, L. D. (2013). Automatic processing of irrelevant co-speech gestures with human but not robot actors. *ACM/IEEE International Conference on Human-Robot Interaction*, 333–340. https://doi.org/10.1109/HRI.2013.6483607

Henkemans, O. A. B., Bierman, B. P., Janssen, J., Neerincx, M. A., Looije, R., van der Bosch, H., & van der Giessen, J. A. (2013). Using a robot to personalise health education for children with diabetes type 1: A pilot study. *Patient Education and Counseling*, *92*(2), 174–181.

Henschel, A., Hortensius, R., & Cross, E. S. (2020). Social cognition in the age of human–robot interaction. *Trends in Neurosciences*, *43*(6), 373–384. https://doi.org/10.1016/j.tins.2020.03.013

Heylen, D., Kopp, S., Marsella, S. C., Pelachaud, C., & Vilhjálmsson, H. (2008). The next step towards a function markup language. *International Workshop on Intelligent Virtual Agents*, 270–280.

Hoetjes, M., Krahmer, E., & Swerts, M. (2015). On what happens in gesture when communication is unsuccessful. *Speech Communication*, *72*, 160–175. https://doi.org/10.1016/j.specom.2015.06.004

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.

Hoffman, G., & Zhao, X. (2020). A primer for conducting experiments in human–robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, *10*(1), 1–31.

Holler, J., & Wilkin, K. (2011). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, *35*(2), 133–153.

Holroyd, A., Rich, C., Sidner, C. L., & Ponsler, B. (2011). Generating connection events for human-robot collaboration. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 241–246. https://doi.org/10.1109/ROMAN.2011.6005245

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

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

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, *15*(3), 495–514.

Hostetter, A. B., & Potthoff, A. L. (2012). Effects of personality and social situation on representational gesture production. *Gesture*, *12*(1), 62–83.

Howley, I., Kanda, T., Hayashi, K., & Rosé, C. (2014). Effects of social presence and social role on help-seeking and learning. *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, 415–422.

Hsieh, W.-F., Sato-Shimokawara, E., & Yamaguchi, T. (2020). Investigation of robot expression style in human-robot interaction. *Journal of Robotics and Mechatronics*, *32*(1), 224–235.

Hua, M., Shi, F., Nan, Y., Wang, K., Chen, H., & Lian, S. (2019). Towards more realistic human-robot conversation: A seq2seq-based body gesture interaction system. *arXiv preprint arXiv:1905.01641*.

Huang, C.-M., & Mutlu, B. (2013). Modeling and evaluating narrative gestures for humanlike robots. *Robotics: Science and Systems*, 57–64.

Huang, C.-M., & Mutlu, B. (2014). Learning-based modeling of multimodal behaviors for humanlike robots. *ACM/IEEE International Conference on Human-Robot Interaction*, 57–64. https://doi.org/10.1145/2559636.2559668

Hwang, E. J., Kyu Ahn, B., Macdonald, B. A., & Seok Ahn, H. (2020). Demonstration of hospital receptionist robot with extended hybrid code network to select responses and gestures. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 8013–8018. https://doi.org/10.1109/ICRA40945.2020.9197160

Iio, T., Shiomi, M., Shinozawa, K., Akimoto, T., Shimohara, K., & Hagita, N. (2011). Investigating entrainment of people's pointing gestures by robot's gestures using a WOZ method. *International Journal of Social Robotics*, *3*(4), 405–414. https://doi.org/10.1007/s12369-011-0112-0

Irfan, B., Kennedy, J., Lemaignan, S., Papadopoulos, F., Senft, E., & Belpaeme, T. (2018). Social psychology and human-robot interaction: An uneasy marriage. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 13–20.

Isaka, T., Aoki, R., Mukawa, N., & Ohshima, N. (2018). Study of socially appropriate robot behaviors in human-robot conversation closure. *ACM International Conference Proceeding Series*, 519–523. https://doi.org/10.1145/3292147.3292243

Ishi, C. T., Machiyashiki, D., Mikata, R., & Ishiguro, H. (2018). A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters*, *3*(4), 3757–3764. https://doi.org/10.1109/LRA.2018.2856281

Ishi, C. T., Mikata, R., & Ishiguro, H. (2020). Person-directed pointing gestures and inter-personal relationship: Expression of politeness to friendliness by android robots. *IEEE Robotics and Automation Letters*, *5*(4), 6081–6088. https://doi.org/10.1109/LRA.2020.3011354

Ishi, C. T., Mikata, R., Minato, T., & Ishiguro, H. (2019). Online processing for speech-driven gesture motion generation in android robots. *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*, 484–490. https://doi.org/10.1109/Humanoids43949.2019.9035066

Ivanov, S. H., Webster, C., & Berezina, K. (2017). Adoption of robots and service automation by tourism and hospitality companies. *Revista Turismo & Desenvolvimento*, *27*(28), 1501–1517.

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, *16*(5), 367–371.

Jain, E., Anthony, L., Aloba, A., Castonguay, A., Cuba, I., Shaw, A., & Woodward, J. (2016). Is the motion of a child perceivably different from the motion of an adult? *ACM Transactions on Applied Perception (TAP)*, *13*(4), 22.

Johanson, D. L., Ahn, H. S., & Broadbent, E. (2020). Improving interactions with healthcare robots: A review of communication behaviours in social and healthcare contexts. *International Journal of Social Robotics*, 1–16.

Jouaiti, M., & Henaff, P. (2019). The sound of actuators: Disturbance in human-robot interactions? *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics, ICDL-EpiRob 2019*, 75–80. https://doi.org/10.1109/DEVLRN.2019.8850697

Jung, H.-W., Seo, Y.-H., Ryoo, M. S., & Yang, H. S. (2004). Affective communication system with multimodality for a humanoid robot, AMI. *4th IEEE/RAS International Conference on Humanoid Robots, 2004.*, *2*, 690–706.

Jutharee, W., & Maneewarn, T. (2016). Gesture reconfiguration from joint failure using genetic algorithm. *International Conference on Control, Automation and Systems*, *0*, 1137–1142. https://doi.org/10.1109/ICCAS.2016.7832455

Kanda, T., Ishiguro, H., Ono, T., Imai, M., & Mase, K. (2002). Multi-robot cooperation for human-robot communication. *Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication.*

Kanda, T., Shimada, M., & Koizumi, S. (2012). Children learning with a social robot. *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 351–358.

Kanero, J., Demir-Lira, E., Koskulu, S., Oranç, C., Franko, I., Küntay, A. C., & Göksun, T. (2018a). How do robot gestures help second language learning? *Earli SIG 5 Abstract book.*

Kanero, J., Geçkin, V., Oranç, C., Mamus, E., Küntay, A. C., & Göksun, T. (2018b). Social robots for early language learning: Current evidence and future directions. *Child Development Perspectives*, *12*(3), 146–151.

Karam, M., & Schraefel, M. (2005). *A taxonomy of gestures in human computer interactions* (Project Report). s.n. https://eprints.soton.ac.uk/261149/

Käser, T., Klingler, S., Schwing, A. G., & Gross, M. (2014). Beyond knowledge tracing: Modeling skill topologies with bayesian networks. *International Conference on Intelligent Tutoring Systems*, 188–198.

Kashii, A., Takashio, K., & Tokuda, H. (2016). Ex-amp robot: Physical avatar for enhancing human to human communication. *HAI 2016 - Proceedings of the 4th International Conference on Human Agent Interaction*, 89–92. https://doi.org/10.1145/2974804.2980509

Kaushik, R., & Simmons, R. (2021). Perception of emotion in torso and arm movements on humanoid robot Quori. *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 62–66. https://doi.org/10.1145/3434074.3447129

Kawaguchi, I., Kodama, Y., Kuzuoka, H., Otsuki, M., & Suzuki, Y. (2016). Effect of embodiment presentation by humanoid robot on social telepresence. *HAI 2016 - Proceedings of the 4th International Conference on Human Agent Interaction*, 253–256. https://doi.org/10.1145/2974804.2980498

Kelly, S. D., Barr, D. J., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language*, *40*(4), 577–592.

Kelly, S. D., Creigh, P., & Bartolotti, J. (2010a). Integrating speech and iconic gestures in a stroop-like task: Evidence for automatic processing. *Journal of Cognitive Neuroscience*, *22*(4), 683–694.

Kelly, S. D., Manning, S. M., & Rodak, S. (2008). Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education. *Language and Linguistics Compass*, *2*(4), 569–588.

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. https://doi.org/10.1080/01690960802365567

Kelly, S. D., Özyürek, A., & Maris, E. (2010b). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260–267.

Kemp, C. C., Edsinger, A., & Torres-Jara, E. (2007). Challenges for robot manipulation in human environments [grand challenges of robotics]. *IEEE Robotics & Automation Magazine*, *14*(1), 20–29.

Kendon, A. (1981). Geography of gesture. *Semiotica*, *37*(1/2), 129–163.

Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of Pragmatics*, *23*(3), 247–279.

Kendon, A. (2004). *Gesture: Visible action as utterance.* Cambridge University Press.

Kennedy, J., Baxter, P., & Belpaeme, T. (2015). The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, (801), 67–74. https://doi.org/10.1145/2696454.2696457

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

Kim, A., Han, J., Jung, Y., & Lee, K. (2013). The effects of familiarity and robot gesture on user acceptance of information. *ACM/IEEE International Conference on Human-Robot Interaction*, 159–160. https://doi.org/10.1109/HRI.2013.6483550

Kim, A., Kum, H., Roh, O., You, S., & Lee, S. (2012). Robot gesture and user acceptance of information in human-robot interaction. *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 279–280. https://doi.org/10.1145/2157689.2157793

Kim, H., Kwak, S. S., & Kim, M. (2008). Personality design of sociable robots by control of gesture design factors. *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 494–499.

Kim, K., Nagendran, A., Bailenson, J. N., Raij, A., Bruder, G., Lee, M., Schubert, R., Yan, X., & Welch, G. F. (2017). A large-scale study of surrogate physicality and gesturing on human–surrogate interactions in a public space. *Frontiers in Robotics and AI*, *4*. https://doi.org/10.3389/frobt.2017.00032

Kipp, M., & Martin, J.-C. (2009). Gesture and emotion: Can basic gestural form features discriminate emotions? *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 1–8.

Kita, S. (2003). *Pointing: Where language, culture, and cognition meet*. Psychology Press.

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, *24*(2), 145–167.

Klahr, D., Triona, L. M., & Williams, C. (2007). Hands on what? the relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. *Journal of Research in Science Teaching*, *44*(1), 183–203.

Ko, W.-R., Lee, J., Jang, M., & Kim, J. (2020). End-to-end learning of social behaviors for humanoid robots. *2020 IEEE International Conference on Systems, Man,* *and Cybernetics (SMC)*, 1200–1205. https://doi.org/10.1109/SMC42975.2020.9283177

Kondo, Y., Takemura, K., Takamatsu, J., & Ogasawara, T. (2012). Planning body gesture of android for multi-person human-robot interaction. *Proceedings - IEEE International Conference on Robotics and Automation*, 3897–3902. https://doi.org/10.1109/ICRA.2012.6224903

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163.

Kory-Westlund, J. M., & Breazeal, C. (2019). A long-term study of young children's rapport, social emulation, and language learning with a peer-like robot playmate in preschool. *Frontiers in Robotics and AI*, *6*, 81. https://doi.org/10.3389/frobt.2019.00081

Kose, H., Akalin, N., Yorganci, R., Ertugrul, B. S., Kivrak, H., Kavak, S., Ozkul, A., Gurpinar, C., Uluer, P., & Ince, G. (2015). ISign: An architecture for humanoid assisted sign language tutoring. In Mohammed, S and Moreno, JC and Kong, K and Amirat, Y (Ed.), *Springer tracts in advanced robotics* (pp. 157–184). https://doi.org/10.1007/978-3-319-12922-8_6

Kose, H., Yorganci, R., Algan, E. H., & Syrdal, D. S. (2012). Evaluation of the robot assisted sign language tutoring using video-based studies. *International Journal of Social Robotics*, *4*(3), 273–283. https://doi.org/10.1007/s12369-012-0142-2

Kraemer, F., Rodriguez, I., Parra, O., Ruiz, T., & Lazkano, E. (2016). Minstrel robots: Body language expression through applause evaluation. *IEEE-RAS International Conference on Humanoid Robots*, 332–337. https://doi.org/10.1109/HUMANOIDS.2016.7803297

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, *57*(3), 396–414.

Kranstedt, A., Kopp, S., & Wachsmuth, I. (2002). Murml: A multimodal utterance representation markup language for conversational agents. *AAMAS'02 Workshop Embodied conversational agents-let's specify and evaluate them!*

Krauss, R. M., & Chen, Y. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and gesture* (pp. 261–283). Cambridge University Press.

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*(1-12), 113–114.

Kucherenko, T., Hasegawa, D., Kaneko, N., Henter, G. E., & Kjellström, H. (2019). On the importance of representations for speech-driven gesture generation. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *4*, 2072–2074.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, *44*(4), 978–990.

Lakin, J. L., & Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, *14*(4), 334–339.

Lalmas, M., O'Brien, H., & Yom-Tov, E. (2014). Measuring user engagement. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *6*(4), 1–132.

Le, Q. A., Hanoune, S., & Pelachaud, C. (2011). Design and implementation of an expressive gesture model for a humanoid robot. *IEEE-RAS International Conference on Humanoid Robots*, 134–140. https://doi.org/10.1109/Humanoids.2011.6100857

Lee, N., Kim, J., Kim, E., & Kwon, O. (2017). The influence of politeness behavior on user compliance with social robots in a healthcare service setting. *International Journal of Social Robotics*, *9*(5), 727–743. https://doi.org/10.1007/s12369-017-0420-0

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*(1), 25–58.

Leite, I., Martinho, C., & Paiva, A. (2013). Social robots for long-term interaction: A survey. *International Journal of Social Robotics*, *5*(2), 291–308.

Leitner, S. (1972). *So lernt man lernen: Der weg zum erfolg [learning to learn: The road to success]*. Freiburg: Herder.

Lemme, A., Freire, A., Barreto, G., & Steil, J. (2013). Kinesthetic teaching of visuomotor coordination for pointing by the humanoid robot iCub. *Neurocomputing*, *112*(SI), 179–188. https://doi.org/10.1016/j.neucom.2012.12.040

Leyzberg, D., Ramachandran, A., & Scassellati, B. (2018). The effect of personalization in longer-term robot tutoring. *ACM Transactions on Human-Robot Interaction (THRI)*, *7*(3), 1–19.

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

Leyzberg, D., Spaulding, S., Toneva, M., & Scassellati, B. (2012). The physical presence of a robot tutor increases cognitive learning gains. *34th Annual Conference of the Cognitive Science Society*, *34*(1), 1882–1887. https://doi.org/ISBN978-0-9768318-8-4

Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, *77*, 23–37.

Li, J., Ju, W., & Nass, C. (2015). Observer perception of dominance and mirroring behavior in human-robot relationships. *ACM/IEEE International Conference on Human-Robot Interaction*, *2015-March*, 133–140. https://doi.org/10.1145/2696454.2696459

Ligthart, M. E., Neerincx, M. A., & Hindriks, K. V. (2020). Design patterns for an interactive storytelling robot to support children's engagement and agency. *ACM/IEEE International Conference on Human-Robot Interaction*, 409–418. https://doi.org/10.1145/3319502.3374826

Ligthart, M. E., van Bindsbergen, K. L., Fernhout, T., Grootenhuis, M. A., Neerincx, M. A., & Hindriks, K. V. (2019). A child and a robot getting acquainted - Interaction design for eliciting self-disclosure. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *1*, 61–70.

Liles, K. R., Perry, C. D., Craig, S. D., & Beer, J. M. (2017). Student perceptions: The test of spatial contiguity and gestures for robot instructors. *ACM/IEEE International Conference on Human-Robot Interaction*, 185–186. https://doi.org/10.1145/3029798.3038297

Lim, A., Ogata, T., & Okuno, H. G. (2011). Converting emotional voice to motion for robot telepresence. *IEEE-RAS International Conference on Humanoid Robots*, 472–479. https://doi.org/10.1109/Humanoids.2011.6100891

Lim, S., Yoon, J., Oh, K., & Cho, S. B. (2009). Gesture based dialogue management using behavior network for flexibility of human robot interaction. *Proceedings of IEEE International Symposium on Computational Intelligence in Robotics and Automation, CIRA*, 592–597. https://doi.org/10.1109/CIRA.2009.5423240

Lindenberg, R., Uhlig, M., Scherfeld, D., Schlaug, G., & Seitz, R. J. (2012). Communication with emblematic gestures: Shared and distinct neural correlates of expression and reception. *Human Brain Mapping*, *33*(4), 812–823.

Liu, P., Glas, D. F., Kanda, T., Ishiguro, H., & Hagita, N. (2017). A model for generating socially-appropriate deictic behaviors towards people. *International Journal of Social Robotics*, *9*(1), 33–49. https://doi.org/10.1007/s12369-016-0348-9

Lohse, M., Rothuis, R., Pérez, J. G., Karreman, D. E., & Evers, V. (2014). Robot gestures make difficult tasks easier: The impact of gestures on perceived workload and task performance. *Conference on Human Factors in Computing Systems - Proceedings*, 1459–1466. https://doi.org/10.1145/2556288.2557274

Louwerse, M. M., & Bangerter, A. (2005). Focusing attention with deictic gestures and linguistic expressions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *27*(27).

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (2010). The Bielefeld speech and gesture alignment corpus (SaGA). In M. Kipp, J.-P. Martin, P. Paggio, & D. Heylen (Eds.), *Lrec 2010 workshop: Multimodal corpora — advances in capturing, coding and analyzing multimodality* (pp. 92–98).

Lun, R., & Zhao, W. (2015). A survey of applications and human motion recognition with Microsoft Kinect. *International Journal of Pattern Recognition and Artificial Intelligence*, *29*(05), 1555008.

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. https://doi.org/10.1002/hbm.21084

Markopoulos, P., Read, J. C., MacFarlane, S., & Hoysniemi, J. (2008). *Evaluating children's interactive products: Principles and practices for interaction designers*. Morgan Kaufmann Publishers Inc.

Marmpena, M., Garcia, F., & Lim, A. (2020). Generating robotic emotional body language of targeted valence and arousal with conditional variational autoencoders. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 357–359. https://doi.org/10.1145/3371382.3378360

Marmpena, M., Lim, A., Dahl, T. S., & Hemion, N. (2019). Generating robotic emotional body language with variational autoencoders. *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 545–551.

Marton, F., & Booth, S. (2013). *Learning and awareness*. Routledge.

Masson-Carro, I., Goudbeek, M., & Krahmer, E. (2015). Coming of age in gesture: A comparative study of gesturing and pantomiming in older children and adults. *Proceedings of the 4th GESPIN — Gesture & Speech in Interaction Conference.*

Masson-Carro, I., Goudbeek, M., & Krahmer, E. (2017). How what we see and what we know influence iconic gesture production. *Journal of Nonverbal Behavior*, *41*(4), 367–394.

Matsui, D., Minato, T., MacDorman, K. F., & Ishiguro, H. (2005). Generating natural motion in an android by mapping human motion. *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3301–3308.

Mavridis, N. (2015). A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, *63*, 22–35.

McGregor, K. K., Rohlfing, K. J., Bean, A., & Marschner, E. (2009). Gesture as a support for word learning: The case of under. *Journal of Child Language*, *36*(4), 807–828. https://doi.org/10.1017/S0305000908009173

McNeil, N. M., Alibali, M. W., & Evans, J. L. (2000). The role of gesture in children's comprehension of spoken language: Now they need it, now they don't. *Journal of Nonverbal Behavior*, *24*(2), 131–150.

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

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* University of Chicago press.

Meena, R., Jokinen, K., & Wilcock, G. (2012). Integration of gestures and speech in human-robot interaction. *3rd IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2012 - Proceedings*, 673–678. https://doi.org/10.1109/CogInfoCom.2012.6421936

Mertens, U. J., & Rohlfing, K. J. (2021). Progressive reduction of iconic gestures contributes to school-aged children's increased word production. *Frontiers in Psychology*, *12*, 1378. https://doi.org/10.3389/fpsyg.2021.651725

Mikata, R., Ishi, C. T., Minato, T., & Ishiguro, H. (2019). Analysis of factors influencing the impression of speaker individuality in android robots. *2019 28th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2019*. https://doi.org/10.1109/RO-MAN46459.2019.8956395

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Mlakar, I., Kacic, Z., & Rojc, M. (2013). TTS-driven synthetic behaviour-generation model for artificial bodies. *International Journal of Advanced Robotic Systems*, *10*. https://doi.org/10.5772/56870

Mlakar, I., Rojc, M., Verdonik, D., & Majhenič, S. (2021). Can turn-taking highlight the nature of non-verbal behavior: A case study. *Types of nonverbal communication*. IntechOpen.

Mohammad, Y., & Nishida, T. (2013). Tackling the correspondence problem. *International Conference on Active Media Technology*, 84–95.

Mori, M. (1970). Bukimi no tani [the uncanny valley]. *Energy*, *7*, 33–35.

Moro, C., Lin, S., Nejat, G., & Mihailidis, A. (2019). Social robots and seniors: A comparative study on the influence of dynamic social features on human–robot interaction. *International Journal of Social Robotics*, *11*(1), 5–24. https://doi.org/10.1007/s12369-018-0488-1

Moshkina, L., Trickett, S., & Trafton, J. G. (2014). Social engagement in public places: A tale of one robot. *ACM/IEEE International Conference on Human-Robot Interaction*, 382–389. https://doi.org/10.1145/2559636.2559678

Mubin, O., Bartneck, C., Feijs, L., Hooft van Huysduynen, H., Hu, J., & Muelver, J. (2012). Improving speech recognition with the robot interaction language. *Disruptive Science and Technology*, *1*(2), 79–88.

Mubin, O., Stevens, C. J., Shahid, S., Mahmud, A. A., & Dong, J.-J. (2013). A review of the applicability of robots in education. *Technology for Education and Learning*, *1*, 209–0015. https://doi.org/10.2316/Journal.209.2013.1.209-0015

Mulder, H., Hoofs, H., Verhagen, J., van der Veen, I., & Leseman, P. P. (2014). Psychometric properties and convergent and predictive validity of an executive function test battery for two-year-olds. *Frontiers in Psychology*, *5*, 733.

Müller, C. (2014). Gestural modes of representation as techniques of depiction. In C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, & J. Bressem (Eds.), *Body–language–communication: An international handbook on multimodality in human interaction* (pp. 1687–1702). De Gruyter Mouton Berlin & Boston.

Muto, Y., Takasugi, S., Yamamoto, T., & Miyake, Y. (2009). Timing control of utterance and gesture in interaction between human and humanoid robot. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 1022–1028. https://doi.org/10.1109/ROMAN.2009.5326319

Nalin, M., Baroni, I., Kruijff-Korbayova, I., Canamero, L., Lewis, M., Beck, A., Cuayahuitl, H., & Sanna, A. (2012). Children's adaptation in multi-session interaction with a humanoid robot. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 351–357. https://doi.org/10.1109/ROMAN.2012.6343778

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453.

Ng-Thow-Hing, V., Luo, P., & Okita, S. (2010). Synchronized gesture and speech production for humanoid robots. *IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings*, 4617–4624. https://doi.org/10.1109/IROS.2010.5654322

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.

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006). Measurement of negative attitudes toward robots. *Interaction Studies*, *7*(3), 437–454. https://doi.org/https://doi.org/10.1075/is.7.3.14nom

Novack, M. A., Goldin-Meadow, S., & Woodward, A. L. (2015). Learning from gesture: How early does it happen? *Cognition*, *142*, 138–147.

O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology*, *59*(6), 938–955.

Okuno, Y., Kanda, T., Imai, M., Ishiguro, H., & Hagita, N. (2008). Providing route directions: Design of robot's utterance, gesture, and timing. *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction, HRI'09*, 53–60. https://doi.org/10.1145/1514095.1514108

Ondáš, S., Juhár, J., Pleva, M., Ferčák, P., & Husovský, R. (2017). Multimodal dialogue system with NAO and VoiceXML dialogue manager. *8th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2017 - Proceedings*, *2018-Janua*. https://doi.org/10.1109/CogInfoCom.2017.8268286

Ondras, J., Celiktutan, O., Bremner, P., & Gunes, H. (2020). Audio-driven robot upper-body motion synthesis. *IEEE Transactions on Cybernetics.*

Ono, T., Imai, M., & Ishiguro, H. (2001). A model of embodied communications with gestures between human and robots. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *23*(23).

Ortega, G., & Özyürek, A. *Generalisable patterns of gesture distinguish semantic categories in communication without language: Evidence from pantomime.* Talk presented at the 7th Conference of the International Society for Gesture Studies (ISGS7). Paris, France. 2016.

Ortega, G., & Özyürek, A. (2020). Systematic mappings between semantic categories and types of iconic representations in the manual modality: A normed database of silent gesture. *Behavior Research Methods*, *52*(1), 51–67.

Özgür, A., Johal, W., Mondada, F., & Dillenbourg, P. (2017). Windfield: Learning wind meteorology with handheld haptic robots. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 156–165.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021a). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/bmj.n71

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021b). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/bmj.n160

Paplu, S. H., Mishra, C., & Berns, K. (2020). Pseudo-randomization in automating robot behaviour during human-robot interaction. *2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 1–6. https://doi.org/10.1109/ICDL-EpiRob48136.2020.9278115

Park, E., Kong, H., Lim, H. T., Lee, J., You, S., & Del Pobil, A. P. (2011). The effect of robot's behavior vs. appearance on communication with humans. *HRI 2011 -*

*Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction*, 219–220. https://doi.org/10.1145/1957656.1957740

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pérez-Mayos, L., Farrús, M., & Adell, J. (2020). Part-of-speech and prosody-based approaches for robot speech and gesture synchronization. *Journal of Intelligent & Robotic Systems*, *99*(2), 277–287. https://doi.org/10.1007/s10846-019-01100-3

Peters, R., Broekens, J., Li, K., & Neerincx, M. A. (2019). Robots Expressing Dominance: Effects of Behaviours and Modulation. *2019 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019*, 461–467. https://doi.org/10.1109/ACII.2019.8925500

Piaget, J., & Cook, M. (1952). *The origins of intelligence in children* (Vol. 8). International Universities Press New York.

Pollmann, K., Ruff, C., Vetter, K., & Zimmermann, G. (2020). Robot vs. voice assistant: Is playing with Pepper more fun than playing with Alexa? *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 395–397. https://doi.org/10.1145/3371382.3378251

Pot, E., Monceaux, J., Gelin, R., & Maisonnier, B. (2009). Choregraphe: A graphical tool for humanoid robot programming. *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 46–51.

Pouw, W., de Wit, J., Bögels, S., Rasenberg, M., Milivojevic, B., & Özyürek, A. (2021). Semantically related gestures move alike: Towards a distributional semantics of gesture kinematics. *Proceedings of the 23rd International Conference on Human-Computer Interaction.*

Pouw, W., & Dixon, J. A. (2020). Gesture networks: Introducing dynamic time warping and network analysis for the kinematic study of gesture ensembles. *Discourse Processes*, *57*(4), 301–319.

Prajod, P., & Hindriks, K. (2020). On the expressivity of a parametric humanoid emotion model. *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 926–931. https://doi.org/10.1109/RO-MAN47096.2020.9223459

Ramachandran, A., Huang, C.-M., Gartland, E., & Scassellati, B. (2018). Thinking aloud with a tutoring robot to enhance learning. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 59–68.

Ramey, A., Gorostiza, J. F., & Salichs, M. A. (2012). A social robot as an aloud reader: Putting together recognition and synthesis of voice and gestures for HRI experimentation. *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 213–214.

Ranatunga, I., Balakrishnan, N., Wijayasinghe, I., & Popa, D. O. (2015). User adaptable tasks for differential teaching with applications to robotic autism therapy. *8th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2015 - Proceedings*. https://doi.org/10.1145/2769493.2775129

Rehm, M., Krogsager, A., & Segato, N. (2016). Perception of affective body movements in hri across age groups: Comparison between results from denmark and japan. *Proceedings - 2015 International Conference on Culture and Computing, Culture and Computing 2015*, 25–32. https://doi.org/10.1109/Culture.and.Computing.2015.14

Repetto, C., Pedroli, E., & Macedonia, M. (2017). Enrichment effects of gestures and pictures on abstract words in a second language. *Frontiers in Psychology*, *8*, 2136.

Riek, L. D., Rabinowitch, T. C., Bremner, P., Pipe, A. G., Fraser, M., & Robinson, P. (2010). Cooperative gestures: Effective signaling for humanoid robots. *5th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2010*, 61–68. https://doi.org/10.1145/1734454.1734474

Robert, L. P., Alahmad, R., Esterwood, C., Kim, S., You, S., & Zhang, Q. (2020). A review of personality in human robot interactions. *arXiv preprint arXiv:2001.11777*.

Robins, B., Dautenhahn, K., Te Boekhorst, R., & Nehaniv, C. L. (2008). Behaviour delay and robot expressiveness in child-robot interactions: A user study on interaction kinesics. *HRI 2008 - Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction: Living with Robots*, 17–24. https://doi.org/10.1145/1349822.1349826

Rodriguez, I., Martínez-Otzeta, J. M., Irigoien, I., & Lazkano, E. (2019). Spontaneous talking gestures using Generative Adversarial Networks. *Robotics and Autonomous Systems*, *114*, 57–65. https://doi.org/10.1016/j.robot.2018.11.024

Rodriguez, I., Martínez-Otzeta, J. M., Lazkano, E., Ruiz, T., & Sierra, B. (2018). On how self-body awareness improves autonomy in social robots. *2017 IEEE International Conference on Robotics and Biomimetics, ROBIO 2017*, *2018-Janua*, 1688–1693. https://doi.org/10.1109/ROBIO.2017.8324661

Roesler, E., Manzey, D., & Onnasch, L. (2021). A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics*, *6*(58), eabj5425.

Rohlfing, K. J. (2019). Learning language from the use of gestures. In J. Horst & J. von Koss Torkildsen (Eds.), *International handbook of language acquisition* (pp. 213–233). Routledge/Taylor & Francis Group.

Rohlfing, K. J., Fritsch, J., Wrede, B., & Jungmann, T. (2006). How can multimodal cues from child-directed interaction reduce learning complexity in robots? *Advanced Robotics*, *20*(10), 1183–1199. https://doi.org/10.1163/156855306778522532

Rohlfing, K. J., Longo, M. R., & Bertenthal, B. I. (2012). Dynamic pointing triggers shifts of visual attention in young infants. *Developmental Science*, *15*(3), 426–435. https://doi.org/10.1111/j.1467-7687.2012.01139.x

Roth, W.-M. (2001). Gestures: Their role in teaching and learning. *Review of Educational Research*, *71*(3), 365–392.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. John Wiley & Sons.

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. https://doi.org/10.1016/j.cedpsych.2012.12.001

Ruffieux, S., Lalanne, D., Mugellini, E., & Abou Khaled, O. (2014). A survey of datasets for human gesture recognition. In M. Kurosu (Ed.), *Human-computer interaction. advanced interaction modalities and techniques* (pp. 337–348). Springer International Publishing.

Sadeghipour, A., Morency, L.-p., & Kopp, S. (2012). Gesture-based object recognition using histograms of guiding strokes. *Proceedings of the British Machine Vision Conference*, 44.1–44.11.

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. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1613–1622.

Sakamoto, D., Kanda, T., Ono, T., Kamashima, M., Imai, M., & Ishiguro, H. (2005). Cooperative embodied communication emerged by interactive humanoid robots. *International Journal of Human Computer Studies*, *62*(2), 247–265. https://doi.org/10.1016/j.ijhcs.2004.11.001

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2013a). To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, *5*(3), 313–323.

Salem, M., Kopp, S., & Joublin, F. (2013b). Generating finely synchronized gesture and speech for humanoid robots: A closed-loop approach. *ACM/IEEE International Conference on Human-Robot Interaction*, 219–220. https://doi.org/10.1109/HRI.2013.6483580

Salem, M., Kopp, S., Wachsmuth, I., Rohlfing, K., & Joublin, F. (2012). Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, *4*(2), 201–217. https://doi.org/10.1007/s12369-011-0124-9

Salem, M., Rohlfing, K., Kopp, S., & Joublin, F. (2011). A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 247–252. https://doi.org/10.1109/ROMAN.2011.6005285

Salvador, M. J., Silver, S., & Mahoor, M. H. (2015). An emotion recognition comparative study of autistic and typically-developing children using the Zeno robot. *Proceedings - IEEE International Conference on Robotics and Automation*, *2015-June*(June), 6128–6133. https://doi.org/10.1109/ICRA.2015.7140059

Saunderson, S., & Nejat, G. (2019). How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics*, *11*(4), 575–608.

Sauppé, A., & Mutlu, B. (2014). Robot deictics: How gesture and context shape referential communication. *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, 342–349.

Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, *12*(1), 13–24.

Scassellati, B., Admoni, H., & Matarić, M. (2012). Robots for use in autism research. *Annual Review of Biomedical Engineering*, *14*, 275–294.

Schlichting, L. (2005). Peabody picture vocabulary test-III-NL. *Amsterdam, the Netherlands: Hartcourt Assessment BV.*

Schodde, T., Bergmann, K., & Kopp, S. (2017). Adaptive robot language tutoring based on bayesian knowledge tracing and predictive decision-making. *Proceedings of ACM/IEEE HRI 2017*, 128–136. https://doi.org/10.1145/2909824.3020222

Schulz, T., Torresen, J., & Herstad, J. (2019). Animation techniques in human-robot interaction user studies: A systematic literature review. *ACM Transactions on Human-Robot Interaction (THRI)*, *8*(2), 1–22.

Sekine, K., Wood, C., & Kita, S. (2018). Gestural depiction of motion events in narrative increases symbolic distance with age. *Language, Interaction and Acquisition*, *9*(1), 40–68.

Seo, J. H., Yang, J. Y., & Kwon, D. S. (2014). Generation of various hand-waving motion of a humanoid robot in a greeting situation. *2014 11th International Conference on Ubiquitous Robots and Ambient Intelligence, URAI 2014*, 374–378. https://doi.org/10.1109/URAI.2014.7057372

Seo, J. H., Yang, J. Y., & Kwon, D. S. (2015). Learning and reproduction of valence-related communicative gesture. *IEEE-RAS International Conference on Humanoid Robots*, *2015-Decem*, 237–242. https://doi.org/10.1109/HUMANOIDS.2015.7363541

Shi, C., Kanda, T., Shimada, M., Yamaoka, F., Ishiguro, H., & Hagita, N. (2010). Easy development of communicative behaviors in social robots. *IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings*, 5302–5309. https://doi.org/10.1109/IROS.2010.5650128

Shimazu, A., Hieida, C., Nagai, T., Nakamura, T., Takeda, Y., Hara, T., Nakagawa, O., & Maeda, T. (2018). Generation of gestures during presentation for humanoid robots. *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 961–968.

Shimpi, P. M., Akhtar, N., & Moore, C. (2013). Toddlers' imitative learning in interactive and observational contexts: The role of age and familiarity of the model. *Journal of Experimental Child Psychology*, *116*(2), 309–323.

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., & Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, *166*(1-2), 140–164.

Silpasuwanchai, C., & Ren, X. (2014). Jump and shoot!: Prioritizing primary and alternative body gestures for intense gameplay. *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, 951–954.

Singer, I., & Gerrits, E. (2015). The effect of playing with tablet games compared with real objects on word learning by toddlers. *Conference Proceedings ICT for Language Learning*, 255–9.

Skantze, G. (2020). Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 101178.

So, W. C., Cheng, C. H., Lam, W. Y., Wong, T., Law, W. W., Huang, Y., Ng, K. C., Tung, H. C., & Wong, W. (2019a). Robot-based play-drama intervention may improve the narrative abilities of Chinese-speaking preschoolers with autism spectrum disorder. *Research in Developmental Disabilities*, *95*. https://doi.org/10.1016/j.ridd.2019.103515

So, W. C., Wong, M. K. Y., Cabibihan, J. J., Lam, C. K. Y., Chan, R. Y. Y., & Qian, H. H. (2016). Using robot animation to promote gestural skills in children with autism spectrum disorders. *Journal of Computer Assisted Learning*, *32*(6), 632–646. https://doi.org/10.1111/jcal.12159

So, W. C., Wong, M. K. Y., Lam, C. K. Y., Lam, W. Y., Chui, A. T. F., Lee, T. L., Ng, H. M., Chan, C. H., & Fok, D. C. W. (2018a). Using a social robot to teach gestural recognition and production in children with autism spectrum disorders. *Disability and Rehabilitation: Assistive Technology*, *13*(6), 527–539. https://doi.org/10.1080/17483107.2017.1344886

So, W. C., Wong, M. K. Y., Lam, W. Y., Cheng, C. H., Ku, S. Y., Lam, K. Y., Huang, Y., & Wong, W. L. (2019b). Who is a better teacher for children with autism? Comparison of learning outcomes between robot-based and human-based interventions in gestural production and recognition. *Research in Developmental Disabilities*, *86*, 62–75. https://doi.org/10.1016/j.ridd.2019.01.002

So, W. C., Wong, M. K. Y., Lam, W. Y., Cheng, C. H., Yang, J. H., Huang, Y., Ng, P., Wong, W. L., Ho, C. L., Yeung, K. L., & Lee, C. C. (2018b). Robot-based intervention may reduce delay in the production of intransitive gestures in Chinese-speaking preschoolers with autism spectrum disorder. *Molecular Autism*, *9*(1). https://doi.org/10.1186/s13229-018-0217-5

Sommers, M. S., & Barcroft, J. (2013). Effects of referent token variability on L2 vocabulary learning. *Language Learning*, *63*(2), 186–210.

Spaulding, S., Gordon, G., & Breazeal, C. (2016). Affect-aware student models for robot tutors. *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 864–872.

St. Clair, A., Mead, R., & Mataric, M. J. (2011). Investigating the effects of visual saliency on deictic gesture production by a humanoid robot. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 210–216. https://doi.org/10.1109/ROMAN.2011.6005266

Stanfield, C., Williamson, R., & Özçalişkan, Ş. (2014). How early do children understand gesture–speech combinations with iconic gestures? *Journal of Child Language*, *41*(2), 462–471.

Stites, L. J., & Özçalışkan, Ş. (2017). Who did what to whom? children track story referents first in gesture. *Journal of Psycholinguistic Research*, *46*(4), 1019–1032.

Stolzenwald, J., & Bremner, P. (2017). Gesture mimicry in social human-robot interaction. *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication*, *2017-Janua*, 430–436. https://doi.org/10.1109/ROMAN.2017.8172338

Suay, H. B., & Chernova, S. (2011). Humanoid robot control using depth camera. *Proceedings of the 6th International Conference on Human-Robot Interaction*, 401–402.

Sugiyama, O., Kanda, T., Imai, M., Ishiguro, H., & Hagita, N. (2007). Natural deictic communication with humanoid robots. *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1441–1448. https://doi.org/10.1109/IROS.2007.4399120

Suhm, B., Myers, B., & Waibel, A. (2001). Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *8*(1), 60–98.

Sunardi, M., & Perkowski, M. (2020). Behavior expressions for social and entertainment robots. *2020 IEEE 50th International Symposium on Multiple-Valued Logic (ISMVL)*, 271–278. https://doi.org/10.1109/ISMVL49045.2020.00058

Szafir, D., & Mutlu, B. (2012). Pay attention! Designing adaptive agents that monitor and improve user engagement. *Conference on Human Factors in Computing Systems - Proceedings*, 11–20. https://doi.org/10.1145/2207676.2207679

Taheri, A., Meghdari, A., & Mahoor, M. H. (2020). A close look at the imitation performance of children with autism and typically developing children using a robotic system. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-020-00704-2

Tahir, Y., Dauwels, J., Thalmann, D., & Magnenat Thalmann, N. (2020). A user study of a humanoid robot as a social mediator for two-person conversations. *International Journal of Social Robotics*, *12*(5), 1031–1044. https://doi.org/10.1007/s12369-018-0478-3

Tanaka, F., Cicourel, A., & Movellan, J. R. (2007). Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences*, *104*(46), 17954–17958.

Tay, J., & Veloso, M. (2012). Modeling and composing gestures for human-robot interaction. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 107–112. https://doi.org/10.1109/ROMAN.2012.6343739

Tellier, M. (2005). How do teacher's gestures help young children in second language acquisition? *International Society of Gesture Studies, ISGS.*

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

Thepsoonthorn, C., Ogawa, K.-i., & Miyake, Y. (2021). The exploration of the uncanny valley from the viewpoint of the robot's nonverbal behaviour. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-020-00726-w

Tielman, M., Neerincx, M., Meyer, J. J., & Looije, R. (2014). Adaptive emotional expression in robot-child interaction. *ACM/IEEE International Conference on Human-Robot Interaction*, 407–414. https://doi.org/10.1145/2559636.2559663

Toh, L. P. E., Causo, A., Tzuo, P.-W., Chen, I.-M., & Yeo, S. H. (2016). A review on the use of robots in education and young children. *Journal of Educational Technology & Society*, *19*(2), 148–163.

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science*, *10*(1), 121–125. https://doi.org/10.1111/j.1467-7687.2007.00573.x

Tomasello, M., Carpenter, M., & Liszkowski, U. (2007). A new look at infant pointing. *Child Development*, *78*(3), 705–722.

Trovato, G., Zecca, M., Sessa, S., Jamone, L., Ham, J., Hashimoto, K., & Takanishi, A. (2013). Towards culture-specific robot customisation: A study on greeting interaction with Egyptians. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 447–452. https://doi.org/10.1109/ROMAN.2013.6628520

Trujillo, J. P., Vaitonyte, J., Simanova, I., & Özyürek, A. (2019). Toward the markerless and automatic analysis of kinematic features: A toolkit for gesture and movement research. *Behavior Research Methods*, *51*(2), 769–777.

Tsiourti, C., Weiss, A., Wac, K., & Vincze, M. (2017). Designing emotionally expressive robots: A comparative study on the perception of communication modalities. *HAI 2017 - Proceedings of the 5th International Conference on Human Agent Interaction*, 213–222. https://doi.org/10.1145/3125739.3125744

Tuyen, N. T. V., Elibol, A., & Chong, N. Y. (2020a). Conditional generative adversarial network for generating communicative robot gestures. *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 201–207. https://doi.org/10.1109/RO-MAN47096.2020.9223498

Tuyen, N. T. V., Elibol, A., & Chong, N. Y. (2020b). Learning from humans to generate communicative gestures for social robots. *2020 17th International Conference on Ubiquitous Robots (UR)*, 284–289. https://doi.org/10.1109/UR49135.2020.9144985

Tuyen, N. T. V., Elibol, A., & Chong, N. Y. (2021). Learning bodily expression of emotion for social robots through human interaction. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(1), 16–30. https://doi.org/10.1109/TCDS.2020.3005907

Valenti, A., Block, A., Chita-Tegmark, M., Gold, M., & Scheutz, M. (2020). Emotion expression in a socially assistive robot for persons with Parkinson's disease. *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. https://doi.org/10.1145/3389189.3389190

Valenzeno, L., Alibali, M. W., & Klatzky, R. (2003). Teachers' gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology*, *28*(2), 187–204.

van de Perre, G., Cao, H.-L., De Beir, A., Esteban, P. G., Lefeber, D., & Vanderborght, B. (2018). Generic method for generating blended gestures and affective functional behaviors for social robots. *Autonomous Robots*, *42*(3), 569–580. https://doi.org/10.1007/s10514-017-9650-0

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. *Educational Psychology Review*, *22*(3), 271–296. https://doi.org/10.1007/s10648-010-9127-6 van den Berghe, R., de Haas, M., Oudgenoeg-Paz, O., Krahmer, E., Verhagen, J., Vogt, P., Willemsen, B., de Wit, J., & Leseman, P. (2021a). A toy or a friend? children's anthropomorphic beliefs about robots and how these relate to second-language word learning. *Journal of Computer Assisted Learning*, *37*(2), 396–410. https://doi.org/10.1111/jcal.12497

van den Berghe, R., Oudgenoeg-Paz, O., Verhagen, J., Brouwer, S., de Haas, M., de Wit, J., Willemsen, B., Vogt, P., Krahmer, E., & Leseman, P. (2021b). Individual differences in children's (language) learning skills moderate effects of robot-assisted second language learning. *Frontiers in Robotics and AI*, *8*, 259. https://doi.org/10.3389/frobt.2021.676248

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

van den Heuvel, W. (2015a). *Bartertown: A single-player human computation game to create a dataset of iconic gestures* [Unpublished manuscript].

van den Heuvel, W. (2015b). *Bartertown: Dataset of iconic gestures.* Media Technology MSc program, Leiden University. www.mediatechnology.leiden.edu/openaccess/bartertown

van Dijk, E. T., Torta, E., & Cuijpers, R. H. (2013). Effects of eye contact and iconic gestures on message retention in human-robot interaction. *International Journal of Social Robotics*, *5*(4), 491–501. https://doi.org/10.1007/s12369-013-0214-y

van Nispen, K., de Sandt-Koenderman, V., Mol, L., Krahmer, E., et al. (2014). Pantomime strategies: On regularities in how people translate mental representations into the gesture modality. *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014)*, 3020–3025.

van Nispen, K., van de Sandt-Koenderman, M., & Krahmer, E. (2018). The comprehensibility of pantomimes produced by people with aphasia. *International Journal of Language & Communication Disorders*, *53*(1), 85–100.

van Nispen, K., van de Sandt-Koenderman, W. M., & Krahmer, E. (2017). Production and comprehension of pantomimes used to depict objects. *Frontiers in Psychology*, *8*, 1095.

van Straten, C. L., Peter, J., & Kühne, R. (2020). Child–robot relationship formation: A narrative review of empirical research. *International Journal of Social Robotics*, *12*(2), 325–344.

Vatavu, R.-D. (2019). The dissimilarity-consensus approach to agreement analysis in gesture elicitation studies. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Venture, G., & Kulić, D. (2019). Robot expressive motions: A survey of generation and evaluation methods. *ACM Transactions on Human-Robot Interaction (THRI)*, *8*(4), 1–17.

Viergutz, A., Flemisch, T., & Dachselt, R. (2014). Increasing the expressivity of humanoid robots with variable gestural expressions. *ACM/IEEE International Conference on Human-Robot Interaction*, 314–315. https://doi.org/10.1145/2559636.2559840

Vilhjálmsson, H., Cantelmo, N., Cassell, J., Chafai, N. E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A. N., Pelachaud, C., et al. (2007). The behavior markup language: Recent developments and challenges. *International Workshop on Intelligent Virtual Agents*, 99–111.

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

Vogt, P., de Haas, M., de Jong, C., Baxter, P., & Krahmer, E. (2017a). Child-robot interactions for second language tutoring to preschool children. *Frontiers in Human Neuroscience*, *11*, 73. https://doi.org/10.3389/fnhum.2017.00073

Vogt, P., Dunk, S., & Poos, P. (2017b). Foreign language tutoring for young adults with severe learning problems. *ACM/IEEE International Conference on Human-Robot Interaction*, 317–318. https://doi.org/10.1145/3029798.3038332

Vogt, P., van den Berghe, R., de Haas, M., Hoffman, L., Kanero, J., Mamus, E., Montanier, J.-M., Oranç, C., Oudgenoeg-Paz, O., García, D. H., et al. (2019). Second language tutoring using social robots: A large-scale study. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 497–505.

Von Laban, R. (1975). *Modern educational dance*. Princeton Book Company Pub.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wakefield, E., Novack, M. A., Congdon, E. L., Franconeri, S., & Goldin-Meadow, S. (2018). Gesture helps learners learn, but not merely by guiding their visual attention. *Developmental Science*, *21*(6), e12664.

Walter, R., Bailly, G., & Müller, J. (2013). StrikeAPose: Revealing mid-air gestures on public displays. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 841–850.

Wang, I., Fraj, M. B., Narayana, P., Patil, D., Mulay, G., Bangar, R., Beveridge, J. R., Draper, B. A., & Ruiz, J. (2017). EGGNOG: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 414–421.

Wang, X., Williams, M. A., Gärdenfors, P., Vitale, J., Abidi, S., Johnston, B., Kuipers, B., & Huang, A. (2014). Directing human attention with pointing. *IEEE RO-MAN 2014 - 23rd IEEE International Symposium on Robot and Human Interactive Communication: Human-Robot Co-Existence: Adaptive Interfaces and Systems for Daily Life, Therapy, Assistance and Socially Engaging Interactions*, 174–179. https://doi.org/10.1109/ROMAN.2014.6926249

Westlund, J. K., Dickens, L., Jeong, S., Harris, P., DeSteno, D., & Breazeal, C. (2015). A comparison of children learning new words from robots, tablets, & people. *Proceedings of the 1st International Conference on Social Robots in Therapy and Education.*

Wicke, P., & Veale, T. (2020). The show must go on: On the use of embodiment, space and gesture in computational storytelling. *New Generation Computing*, *38*(4, SI), 565–592. https://doi.org/10.1007/s00354-020-00106-y

Willemsen, B., de Wit, J., Krahmer, E., de Haas, M., & Vogt, P. (2018). Context-sensitive natural language generation for robot-assisted second language tutoring. *Proceedings of the Workshop on NLG for Human–Robot Interaction.*

Wobbrock, J. O., Morris, M. R., & Wilson, A. D. (2009). User-defined gestures for surface computing. *Proceedings of the 2009 CHI Conference on Human Factors in Computing Systems*, 1083–1092.

Wolfert, P., Robinson, N., & Belpaeme, T. (2021). A review of evaluation practices of gesture generation in embodied conversational agents. *arXiv preprint arXiv:2101.03769.*

Wu, Y., Wang, R., Tay, Y. L., & Wong, C. J. (2017). Investigation on the roles of human and robot in collaborative storytelling. *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).*

Wykowska, A., Kajopoulos, J., Obando-Leitón, M., Chauhan, S. S., Cabibihan, J. J., & Cheng, G. (2015). Humans are well tuned to detecting agents among non-agents: Examining the sensitivity of human perception to behavioral characteristics of intentional systems. *International Journal of Social Robotics*, *7*(5), 767–781. https://doi.org/10.1007/s12369-015-0299-6

Xu, J., Broekens, J., Hindriks, K., & Neerincx, M. A. (2013). Mood expression through parameterized functional behavior of robots. *22nd IEEE International Symposium on Robot and Human Interactive Communication:" Living Together, Enjoying Together, and Working Together with Robots!", IEEE RO-MAN 2013*, 533–540.

Xu, J., Broekens, J., Hindriks, K., & Neerincx, M. A. (2014). Effects of bodily mood expression of a robotic teacher on students. *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2614–2620.

Xu, J., Broekens, J., Hindriks, K., & Neerincx, M. A. (2015a). Effects of a robotic storyteller's moody gestures on storytelling perception. *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, 449–455. https://doi.org/10.1109/ACII.2015.7344609

Xu, J., Broekens, J., Hindriks, K., & Neerincx, M. A. (2015b). Mood contagion of robot body language in human robot interaction. *Autonomous Agents and Multi-Agent Systems*, *29*(6), 1216–1248. https://doi.org/10.1007/s10458-015-9307-3

Xu, K. (2019). First encounter with robot Alpha: How individual differences interact with vocal and kinetic cues in users' social responses. *New Media and Society*, *21*(11-12), 2522–2547. https://doi.org/10.1177/1461444819851479

Yadollahi, E., Johal, W., Paiva, A., & Dillenbourg, P. (2018). When deictic gestures in a robot can harm child-robot collaboration. *IDC 2018 - Proceedings of the 2018 ACM Conference on Interaction Design and Children*, 195–206. https://doi.org/10.1145/3202185.3202743

Yamazaki, K., Yamazaki, A., Ikeda, K., Liu, C., Fukushima, M., Kobayashi, Y., & Kuno, Y. (2016). "i'll be there next": A multiplex care robot system that conveys service order using gaze gestures. *ACM Transactions on Interactive Intelligent Systems*, *5*(4). https://doi.org/10.1145/2844542

Yang, G.-Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., Jacobstein, N., Kumar, V., McNutt, M., Merrifield, R., et al. (2018). The grand challenges of science robotics. *Science Robotics*, *3*(14), eaar7650.

Yoon, Y., Ko, W.-R., Jang, M., Lee, J., Kim, J., & Lee, G. (2019). Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. *2019 International Conference on Robotics and Automation (ICRA)*, 4303–4309.

Yu, C., & Tapus, A. (2020). SRG3: Speech-driven robot gesture generation with GAN. *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 759–766. https://doi.org/10.1109/ICARCV50220.2020.9305330

Zabala, U., Rodriguez, I., Martínez-Otzeta, J. M., Irigoien, I., & Lazkano, E. (2021). Quantitative analysis of robot gesticulation behavior. *Autonomous Robots*, *45*(1), 175–189.

Zaga, C., Lohse, M., Truong, K. P., & Evers, V. (2015). The effect of a robot's social character on children's task engagement: Peer versus tutor. *International Conference on Social Robotics*, 704–713.

Zhang, P., & de Haas, M. (2020). Effects of pitch gestures on learning chinese orthography with a social robot. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 180–182. https://doi.org/10.1145/3371382.3378350

Zhao, X., Cusimano, C., & Malle, B. F. (2016). Do people spontaneously take a robot's visual perspective? *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, 335–342.

Zhao, X., & Malle, B. F. (2019). Seeing through a robot's eyes: Spontaneous perspective taking toward humanlike machines. https://doi.org/10.31234/osf.io/8z72c

Zheng, M., Liu, P. X., & Meng, M. Q. (2019). Interpretation of human and robot emblematic gestures: Howdo they differ? *International Journal of Robotics and Automation*, *34*(1), 55–70. https://doi.org/10.2316/J.2019.206-5163

Zheng, M., & Meng, M. Q. (2012). Designing gestures with semantic meanings for humanoid robot. *2012 IEEE International Conference on Robotics and Biomimetics, ROBIO 2012 - Conference Digest*, 287–292. https://doi.org/10.1109/ROBIO.2012.6490981

Zheng, Z., Young, E. M., Swanson, A. R., Weitlauf, A. S., Warren, Z. E., & Sarkar, N. (2016). Robot-mediated imitation skill training for children with autism. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *24*(6), 682–691. https://doi.org/10.1109/TNSRE.2015.2475724

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*.

Żywiczyński, P., Wacewicz, S., & Orzechowski, S. (2017). Adaptors and the turn-taking mechanism: The distribution of adaptors relative to turn borders in dyadic conversation. *Interaction Studies*, *18*(2), 276–298.

# Summary

In recent years we have seen a number of technological innovations in (primary school) classrooms. It is important to explore the role that technology can play in this context, because classroom sizes are increasing and, as a result, it is becoming increasingly more challenging for teachers to accommodate their students' individual needs. Social robots could be considered the next iteration of technology that could potentially support — but certainly not replace — teachers, for example by offering individual tutoring sessions to students. Compared to existing technologies that are already frequently being used in education, such as tablet devices, robots have the added benefit of being physically present in the context where learning takes place. This allows them to also connect with the learner socially, for example by using natural speech to talk, and by using non-verbal modes of communication such as hand gestures. This social component is known to play an important role when learning from other people, therefore we set out to investigate whether this also applies to educational interactions between children and a robot. In this thesis we focus specifically on the use of hand gestures, which can be considered a defining property of social robots, to support second language learning.

The research presented in this thesis was conducted as part of the L2TOR project, in which we studied whether the SoftBank Robotics NAO robot could successfully be used to teach second language vocabulary to children of 4–6 years old. Research has shown that learning a second language, especially at an early age, provides several benefits for the future, for example in terms of career prospects. We developed a number of tutoring interactions, consisting of the robot and the child together playing games or going through scenarios on a tablet device. These interactions were put to the test in several experiments at primary schools in the Netherlands, where children tried to learn English vocabulary with the robot. We measured whether the robot was indeed capable of helping children learn second language vocabulary, whether the children were engaged with the educational task and with the robot, and how they perceived the robot (e.g., as more of a human or more of a 'thing'). This thesis investigates the role of the robot's use of hand gestures in improving language learning performance and engagement.

In **Chapter 2**, we provide an overview of the state of research into robot-performed gestures. We focus on social (human-looking) robots, but we do examine a number of domains other than education where these robots are being used, such as the hospitality industry and healthcare. Our overview shows that gestures can be used to change the way the robot is perceived (e.g., its personality or mood), they can

result in more engagement with and enjoyment of the interaction, they can improve performance on human-robot collaborative tasks (including educational tasks), and they can support the robot's interactions with people with special needs, such as autistic children. We have further identified a number of outstanding questions in the field of robot-performed gesture research that can help pave the way for future research, such as a need for more studies into mirroring or reenactment of the robot's gestures. Several of these outstanding questions are addressed in the later chapters of this thesis.

**Chapters 3, 4, and 6** describe three experimental studies that were conducted to investigate whether robot-performed gestures can indeed result in greater learning outcomes, in terms of more English vocabulary words learned, as well as greater levels of engagement (in Chapters 3 and 6). Chapter 4 describes a longitudinal study consisting of seven sessions with the robot, while Chapters 3 and 6 were single-session studies. All three studies confirm that gestures can be used to facilitate second language vocabulary learning, although we found that this particularly applied to the older children in our studies (those that were approximately six years old). Furthermore, the robot's use of gestures was shown to result in higher levels of engagement, particularly with the robot.

In **Chapter 5** we introduce a dataset of human-performed gestures, that was collected by having the robot play a game of charades with visitors to the NEMO science museum in Amsterdam, and the Lowlands music festival in Biddinghuizen. We observed, in line with existing gesture research, that people tend to vary in the types of gestures they perform for certain concepts. For example, a pencil can be depicted by 'becoming' the pencil and raising one's hands in a pointy shape above the head, or by depicting the act of writing or drawing something on a piece of paper. The gesture recordings that were collected can be used as guidelines to design gestures for a robot or virtual agent, making use of what we have learned from human gesturing behavior. Because we observed variation in gesturing behavior by people that participated in our data collection study, in Chapter 6 we had the robot vary its gesturing behavior as well. However, we but did not find an effect on children's learning outcomes or levels of engagement. More research is needed to further explore the effects of gesture variation, as this has also not yet been studied in human-human communication settings. We are particularly interested to see how variation may affect long-term engagement across multiple interactions with the robot, as well as the way the robot is perceived — two important aspects of human-robot interaction that were not included in the current study.

## Implications and conclusion

We have studied the potential of using social robots as second language tutors for children, and specifically focused on the role that the robot's use of hand gestures could play. We found that social robots are able to support second language vocabulary training, and that particularly the older children in our studies (of approximately six years old) further benefited from the robot's use of hand gestures. The robot's use of gestures also resulted in higher levels of engagement, which could potentially maintain children's interest in the robot for a longer period of time. This thesis thereby furthers our understanding of how robots can contribute to second language tutoring, and how to make use of their physical presence in the context where learning takes place by means of hand gestures. The source code of our experiments, as well as the dataset of recorded gestures from Chapter 5, have been made publicly available to support future research.

# Acknowledgements

“It’s not what you know, but who you know, y’know?”
- Tifa Lockhart

Dit proefschrift had nooit tot stand kunnen komen zonder de beste mentoren die een promovendus zich kan wensen: Emiel Krahmer en Paul Vogt. Emiel, bedankt dat je me toen hebt ontvangen in Tilburg, voor een gesprek wat ‘echt geen sollicitatie was, gewoon een informeel kennismakingsgesprek’ ;-) Vanaf dat moment heb je mij altijd op mijn gemak kunnen stellen, met veel vrijheid om mijn eigen pad te volgen en om de verantwoordelijkheid te nemen, zowel bij het proefschrift als bij het begeleiden van scriptiestudenten. Bedankt voor alle steun, zowel inhoudelijk als persoonlijk! Ik heb ontzettend veel bewondering voor het werk dat je doet — het begeleiden van een groot aantal promovendi, en ook nog de hele afdeling draaiende houden. Paul, wat waren onze afspraken altijd heel gezellig en leerzaam tegelijkertijd! Jouw gedetailleerde opmerkingen hebben mijn onderzoek en schrijven naar een hoger niveau getild. Heel concreet zal jouw tip om te beginnen met inzichtelijke grafieken en daarna pas ingewikkelde statistiek te presenteren me altijd bij blijven, en dit draag ik nu ook regelmatig over op scriptiestudenten. Ontzettend bedankt in het bijzonder voor je hulp bij de laatste loodjes van mijn proefschrift, ook al was je inmiddels al volop aan het werk bij Hanze Hogeschool. Ze hebben het daar enorm getroffen met jou, en ik hoop dat we toch snel weer een samenwerking aan kunnen gaan! Emiel en Paul, ik vond het ook heel leuk om samen conferenties af te gaan en robots zoveel mogelijk op de kaart te zetten. Mijn mooiste herinnering is denk ik toch wel ons uitstapje naar Lowlands, waar ik van Paul ’s ochtends op de camping een lekkere kop koffie kreeg, en ik met Emiel op platenjacht ging.

Team D329, Mirjam en Bram, bedankt voor alle gezelligheid en inspiratie tijdens onze tijd samen op kantoor. Ik herinner me dat er dagen waren waarop we het rustigaan deden, en dagen waarop we tot in de avond aan het doorzetten waren om de robot mee te laten werken. Dit zijn allebei (nu) toffe herinneringen geworden! Zonder jullie was het L2TOR project heel veel saaier geweest, bedankt dat jullie het tot zo’n succesvolle maar ook plezierige ervaring hebben gemaakt! Toch een heel grappig idee dat we straks alledrie als doctor door het leven mogen gaan :-)

This brings me to the amazing L2TOR project group: Thank you for having me as part of this wonderful project, and thank you for all the fun and productive times during the project. A special thanks to “team Chicago” (Rianne, Laura, Thorsten) — I won’t forget our little cocktail adventure at the top of the tower, followed by the cocktail adventure at the Sugar Factory.

Thank you to the members of my PhD committee: Tony Belpaeme, Maartje de Graaf, Spencer Kelly, Panos Markopoulos, Katharina Rohlfing, and Marc Swerts: Your comments on the thesis were incredibly valuable. They have been addressed in the version you see before you now, and I will also incorporate your feedback into the two papers that are currently under review. I was lucky to have met most of you before already, and I hope that we can continue meeting each other and collaborating in the future! Thank you also to my paranymphs, Mirjam and Tudor (and bonus Debargha), for having my back before and during the defense. Marjolijn, heel leuk dat je zo graag de voorzitter wilde zijn!

Er zijn een aantal mensen die hebben bijgedragen aan het tot stand komen van verschillende hoofdstukken, zonder wie dit proefschrift waarschijnlijk nog lang niet af was geweest. Natuurlijk wil ik alle kinderen die aan het onderzoek hebben meegedaan, hun ouders, en de scholen die ons welkom heetten enorm bedanken. Hetzelfde geldt voor het NEMO wetenschapsmuseum en het Lowlands muziekfestival, waar we ons enorm thuis voelden terwijl we daar onderzoek mochten doen. Daarnaast wil ik Martijn Faes, Laurette Gerts, Sanne van Gulik, Annabella Hermans, Esmee Kramer, Madée Kruijt, Marije Merckens, David Mogendorff, Sam Muntjewerf, Reinjet Oostdijk, Marijn Peters Rit, Laura Pijpers, Emmy Rintjema, Chani Savelberg, Robin Sonders, Sirkka van Straalen, Sabine Verdult, Esmee Verheem, Pieter Wolfert, Hugo Zijlstra en Michelle Zomers bedanken voor hun hulp bij het uitvoeren van de experimenten, en Henrike Colijn voor het helpen bij het annoteren van de betrokkenheid van de kinderen in Hoofdstuk 6. Peggy, enorm bedankt voor de leuke samenwerking, en specifiek voor je hulp bij de evaluaties van de robotgebaren. Thank you to Chrissy Cook and Elske van der Vaart for lending us their voice for spoken content used in various post-tests and tablet interactions. David Peeters, beste kamergenoot ooit, bedankt voor jouw feedback op een eerste versie van Hoofdstuk 5!

Vloeiende overgang van mijn kamergenoot naar Tilburg University. Bedankt, beste collega's: Ik heb me vanaf het begin heel welkom en op mijn gemak gevoeld bij jullie. Als ik alle toffe collega's zou moeten noemen loop ik het risico dat ik iemand vergeet, maar weet dat ik jullie allemaal enorm waardeer. Een bijzondere shout out naar het gezelligheidsclubje, en het schrijfweekend-clubje (wat volgens mij nooit echt een weekend is geweest, en waar meestal niet echt veel geschreven wordt). Eva en Lauraine, enorm bedankt dat jullie achter de schermen altijd alles zo goed voor ons weten te regelen!

Fons en nogmaals Marjolijn en Emiel, ik kan jullie niet genoeg bedanken voor alle mogelijkheden die jullie mij hebben geboden bij New Media Design. Het feit dat ik al zo vroeg de kans kreeg om onderwijservaring op te doen heeft me enorm geholpen, en het heeft me ook gemotiveerd om door te zetten met het afronden van mijn promotietraject. Ik wil ook alle studenten aan onze opleiding bedanken die ik tot nu toe de oren van hun hoofd heb mogen kletsen: Jullie hebben mij (misschien zonder het te weten) enorm geïnspireerd, en ik ben nog lang niet uitgekletst! In het bijzonder wil ik Arold Brandse bedanken voor zijn hulp bij de studie beschreven in Hoofdstuk 6 van dit proefschrift, en Renée de Leau voor het ontwerpen van de mooiste kaft en hoofdstuktitelpagina's die ik me kan voorstellen.

Naast mijn waardevolle werkfamilie is er natuurlijk ook mijn 'gewone' familie, die er ook voor heeft gezorgd dat ik mijn doel niet uit het oog verloor. Ook hier geldt dat ik geen lijstjes durf te noemen omdat ik bang ben dat ik iemand vergeet, maar in het bijzonder wil ik toch even Marion, Marcus en Marit noemen omdat ik daar altijd terecht kon (en aan kon schuiven voor eten!) tijdens het schrijven van het proefschrift, en nu natuurlijk nog steeds. Erg leuk ook dat jullie in NEMO langskwamen!

Ik heb ook vrienden, jawel! Luuk, Liza, Ferdi, Marloes, Bram en Saskia: bedankt voor de gezelligheid en de steun de afgelopen jaren, dat we nog maar veel leuke dingen mogen doen! Laurie, Dirk, Martijn, Colinde, Stefanie, Ron, Yvonne en Frank (+ kids natuurlijk, de nieuwe generatie scheetjes): jullie zijn mijn oudste vrienden, bedankt voor alle steun en inspiratie, al zoveel jaren! Blij dat jullie niet doorhebben dat ik de pubquiz steeds expres verlies zodat ik hem niet hoef te organiseren ;-) Avans/UU crew: Bas, Bert en Erik, wat zijn we groot geworden hè ;-) Bedankt voor alle leuke tijden tot nu toe! Erik, Maike, Fae en Lily (en verdere familie!) ook in het bijzonder bedankt dat ik altijd welkom ben bij jullie, en voor alle gezellige late night gesprekken. Ik ben een heel trotse peetoom :-) Finally (I hope I didn't miss anyone), thank you to my TU Eindhoven crew, for all the fun times and all the things I've learned from you. Look at that, I made it to the finish line! A special thanks to Debargha, Kate, Roxana, Pedro, and Kalyani for the great times, not to mention the amazing food, in Eindhoven! Sanne, tofste buurvrouw, bedankt voor alle gezellige wandelingen en gesprekken!

Lieve Susanne, mijn bestie, bedankt dat je de laatste, waarschijnlijk meest intensieve maanden van deze reis zo leuk hebt gemaakt. Jouw rol hierin is veel groter dan je waarschijnlijk denkt en ik ben blij dat ik jou heb leren kennen. Bedankt dat je er op een of andere manier in slaagt om mijn stress als sneeuw voor de zon te laten verdwijnen. We hebben allerlei leuke dingen samen gedaan, en dankzij jou ben ik de leukste versie van mezelf — en dat is 220 :-) PS: Véronique, bedankt voor de ketchupchipstip ;-)

Mam, ik had dit nooit kunnen doen zonder jou. Bedankt voor alle liefde, steun, humor, en plaatsvervangende stress. Je bent m'n beste vriendin, en ik draag dit werk met liefde aan jou op. Pap, bedankt voor jouw liefde en inspiratie. Ik mis je nog iedere dag, en zal nooit vergeten wat je voor ons betekent. Zo vraag ik me nog regelmatig af: wat zou senior doen? En dan kom ik er, met jouw hulp, altijd weer uit. Volgens mij had je het meteen geloofd als ik je 15 jaar geleden zou hebben gezegd dat ik nu betaald zou worden om met robots en videogames te spelen. Ik draag met trots onze naam, en schrijf deze met plezier boven de artikelen die ik publiceer. Het voelt dan toch een beetje alsof we het samen hebben gedaan, zoals hoe we vroeger ook samen allerlei leuke projecten ondernamen. Gefeliciteerd met je verjaardag, ik hoop dat mijn verdediging een leuk verjaardagscadeau is!

# List of Publications

## Journal publications

**de Wit**, J., Vogt, P., & Krahmer, E. (n.d.-a). The design and observed effects of robot-performed manual gestures: A systematic review [Submitted for journal publication].

**de Wit**, J., Willemsen, B., de Haas, M., van den Berghe, R., Leseman, P., Oudgenoeg-Paz, O., Verhagen, J., Vogt, P., & Krahmer, E. (n.d.-b). Designing and evaluating iconic gestures for child-robot second language learning [Submitted for journal publication].

Geerts, J., **de Wit**, J., & de Rooij, A. (2021). Brainstorming with a social robot facilitator: Better than human facilitation due to reduced evaluation apprehension? *Frontiers in Robotics and AI*, *8*, 156. https://doi.org/10.3389/frobt.2021.657291

Leeuwestein, H., Barking, M., Sodacı, H., Oudgenoeg-Paz, O., Verhagen, J., Vogt, P., Aarts, R., Spit, S., de Haas, M., **de Wit**, J., & Leseman, P. (2021). Teaching Turkish-Dutch kindergartners dutch vocabulary with a social robot: Does the robot's use of Turkish translations benefit children's Dutch vocabulary learning? *Journal of Computer Assisted Learning*, *37*(3), 603–620. https://doi.org/https://doi.org/10.1111/jcal.12510

van den Berghe, R., de Haas, M., Oudgenoeg-Paz, O., Krahmer, E., Verhagen, J., Vogt, P., Willemsen, B., **de Wit**, J., & Leseman, P. (2021a). A toy or a friend? Children's anthropomorphic beliefs about robots and how these relate to second-language word learning. *Journal of Computer Assisted Learning*, *37*(2), 396–410. https://doi.org/https://doi.org/10.1111/jcal.12497

van den Berghe, R., Oudgenoeg-Paz, O., Verhagen, J., Brouwer, S., de Haas, M., **de Wit**, J., Willemsen, B., Vogt, P., Krahmer, E., & Leseman, P. (2021b). Individual differences in children's (language) learning skills moderate effects of robot-assisted second language learning. *Frontiers in Robotics and AI*, *8*, 259. https://doi.org/10.3389/frobt.2021.676248

**de Wit**, J., Krahmer, E., & Vogt, P. (2020a). Introducing the NEMO-Lowlands iconic gesture dataset, collected through a gameful human–robot interaction. *Behavior Research Methods*, 1–18.

**de Wit**, J., van der Kraan, A., & Theeuwes, J. (2020b). Live streams on Twitch help viewers cope with difficult periods in life. *Frontiers in Psychology*, *11*, 3162. https://doi.org/10.3389/fpsyg.2020.586975

Quaedackers, L., **de Wit**, J., Pillen, S., Van Gilst, M., Batalas, N., Lammers, G. J., Markopoulos, P., & Overeem, S. (2020). A mobile app for longterm monitoring of narcolepsy symptoms: Design, development, and evaluation. *JMIR mHealth and uHealth*, *8*(1), e14939.

Belpaeme, T., Vogt, P., van den Berghe, R., Bergmann, K., Göksun, T., de Haas, M., Kanero, J., Kennedy, J., Küntay, A. C., Oudgenoeg-Paz, O., Papadopoulos, F., Schodde, T., Verhagen, J., Wallbridge, C. D., Willemsen, B., **de Wit**, J., Geçkin, V., Hoffmann, L., Kopp, S., … Pandey, A. K. (2018). Guidelines for designing social robots as second language tutors. *International Journal of Social Robotics*, *10*(3), 325–341.

Soute, I., Vacaretu, T., **de Wit**, J., & Markopoulos, P. (2017). Design and evaluation of RaPIDO, a platform for rapid prototyping of interactive outdoor games. *ACM Trans. Comput.-Hum. Interact.*, *24*(4). https://doi.org/10.1145/3105704

## Papers in conference proceedings (peer reviewed)

Pouw, W., **de Wit**, J., Bögels, S., Rasenberg, M., Milivojevic, B., & Özyürek, A. (2021). Semantically related gestures move alike: Towards a distributional semantics of gesture kinematics. *Proceedings of the 23rd International Conference on Human-Computer Interaction.*

**de Wit**, J. (2021). A unified model of game design, through the lens of user experience. *Extended abstracts of the 2021 CHI conference on human factors in computing systems.* Association for Computing Machinery. https://doi.org/10.1145/3411763.3451778

Antheunis, M., Croes, E., van der Lee, C., & **de Wit**, J. (2020). Your secret is safe with me. The willingness to disclose intimate information to a chatbot and its impact on emotional well-being. *Etmaal van de Communicatiewetenschap.*

**de Wit**, J., Brandse, A., Krahmer, E., & Vogt, P. (2020). Varied human-like gestures for social robots: Investigating the effects on children's engagement and language learning. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 359–367. https://doi.org/10.1145/3319502.3374815

van Minkelen, P., Gruson, C., van Hees, P., Willems, M., **de Wit**, J., Aarts, R., Denissen, J., & Vogt, P. (2020). Using self-determination theory in social robots to increase motivation in L2 word learning. *Proceedings of the 2020 ACM/IEEE*

*International Conference on Human-Robot Interaction*, 369–377. https://doi.org/10.1145/3319502.3374828

**de Wit**, J., Willemsen, B., de Haas, M., Krahmer, E., Vogt, P., Merckens, M., Oostdijk, R., Savelberg, C., Verdult, S., & Wolfert, P. (2019). Playing charades with a robot: Collecting a large dataset of human gestures through HRI. *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, 634–635.

Johal, W., Sandygulova, A., **de Wit**, J., de Haas, M., & Scassellati, B. (2019). Robots for learning - R4L: Adaptive learning. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 693–694. https://doi.org/10.1109/HRI.2019.8673109

Vogt, P., van den Berghe, R., de Haas, M., Hoffman, L., Kanero, J., Mamus, E., Montanier, J.-M., Oranç, C., Oudgenoeg-Paz, O., García, D. H., Papadopoulos, F., Schodde, T., Verhagen, J., Wallbridge, C. D., Willemsen, B., **de Wit**, J., Belpaeme, T., Göksun, T., Kopp, S., … Pandey, A. K. (2019a). Second language tutoring using social robots: A large-scale study. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 497–505.

Vogt, P., van den Berghe, R., de Haas, M., Hoffman, L., Kanero, J., Mamus, E., Montanier, J.-M., Oranç, C., Oudgenoeg-Paz, O., García, D. H., Papadopoulos, F., Schodde, T., Verhagen, J, Wallbridge, C. D., Willemsen, B., **de Wit**, J., Belpaeme, T., Göksun, T., Kopp, S., … Pandey, A. K. (2019b). Second language tutoring using social robots: L2TOR - the movie. *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, 373.

**de Wit**, J., Schodde, T., Willemsen, B., Bergmann, K., de Haas, M., Kopp, S., Krahmer, E., & Vogt, P. (2018). The effect of a robot's gestures and adaptive tutoring on children's acquisition of second language vocabularies. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 50–58. https://doi.org/10.1145/3171221.3171277

Leeuwestein, H., Barking, M., Sodacı, H., Aarts, R., **de Wit**, J., Oudgenoeg-Paz, O., Verhagen, J., & Vogt, P. (2018). Bilingual robots teaching L2 vocabulary to immigrant children. *EARLI SIG 5 Conference 2018: ECEC 2.0: Future Challenges for Early Childhood Education and Care.*

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. *Companion of the 2018 ACM/IEEE*

*International Conference on Human-Robot Interaction*, 219–220. https://doi.org/10.1145/3173386.3177059

van den Berghe, R., Oudgenoeg-Paz, O., Verhagen, J., de Haas, M., **de Wit**, J., & Willemsen, B. (2018). Investigating the effectiveness of a social robot for supporting children's L2 learning. *Conference on Multilingualism.*

Manojlovic, S., Gavrilo, K., **de Wit**, J., Khan, V.-J., & Markopoulos, P. (2016). Exploring the potential of children in crowdsourcing. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 1250–1256. https://doi.org/10.1145/2851581.2892312

## Workshop papers (peer reviewed)

**de Wit**, J., Krahmer, E., & Vogt, P. (2019a). Social robots as language tutors: Challenges and opportunities. *Proceedings of the Workshop on the Challenges of Working on Social Robots that Collaborate with People, ACMCHI Conference on Human Factors in Computing Systems (CHI2019 SIRCHI Workshop).*

**de Wit**, J., Pijpers, L., van den Berghe, R., Krahmer, E., & Vogt, P. (2019b). Why UX research matters for HRI: The case of tablets as mediators. *Proceedings of the Workshop on the Challenges of Working on Social Robots that Collaborate with People, ACMCHI Conference on Human Factors in Computing Systems (CHI2019 SIRCHI Workshop).*

Slegers, K., de Rooij, A., van Enschot, R., Elloumi, L., van der Laan, N., & **de Wit**, J. (2019). Academic challenges in HCI education–the New Media Design bachelor and master programs. *Proceedings of the EDUCHI 2019 workshop, ACMCHI Conference on Human Factors in Computing Systems (CHI2019).*

van der Lee, C., Croes, E., **de Wit**, J., & Antheunis, M. (2019). Digital confessions: Exploring the role of chatbots in self-disclosure. *CONVERSATIONS 2019*, *7*, 21.

**de Wit**, J., Willemsen, B., de Haas, M., Wolfert, P., Vogt, P., & Krahmer, E. (2018). Playful exploration of a robot's gesture production and recognition abilities. *Workshop on Gesture & Technology, Warwick.*

Kallergi, A., **de Wit**, J., Bálint, K., de Rooij, A., Krahmer, E., & Maes, F. (2018). HCI education beyond HCI studies: Insights from the New Media Design program. *Proceedings of the EDUCHI 2018 workshop, ACMCHI Conference on Human Factors in Computing Systems (CHI2018).*

Vogt, P., **de Wit**, J., Schodde, T., Willemsen, B., Bergmann, K., de Haas, M., Kopp, S., & Krahmer, E. (2018a). Iconic gestures improve second language learning from a social robot. *ISGS: International Society for Gesture Studies, Cape Town, South Africa.*

Vogt, P., Willemsen, B., **de Wit**, J., de Haas, M., & Krahmer, E. (2018b). Personalized and multimodal interactions for second language tutoring using a social robot. *Symposium "Social Robots for Language Learning" at EARLI SIG 5 Conference 2018.*

Willemsen, B., **de Wit**, J., Krahmer, E., de Haas, M., & Vogt, P. (2018). Context-sensitive natural language generation for robot-assisted second language tutoring. *Proceedings of the Workshop on NLG for Human–Robot Interaction*, 1–7.

**de Wit**, J., Schodde, T., Willemsen, B., Bergmann, K., de Haas, M., Kopp, S., Krahmer, E., & Vogt, P. (2017). Exploring the effect of gestures and adaptive tutoring on children's comprehension of L2 vocabularies. *Proceedings of the Workshop R4L at ACM/IEEE Human-Robot Interaction 2017.*

Vogt, P., **de Wit**, J., Schodde, T., Willemsen, B., Bergmann, K., de Haas, M., Kopp, S., & Krahmer, E. (2017). Adaptation and gestures in second language tutoring using social robots. *Workshop on Early Literacy and (Digital) Media.*

**de Wit**, J., Manojlovic, S., Gavrilo, K., & Khan, V.-J. (2016). Crowdsourcing for children: Exploring threats and opportunities. *International reports on socio-informatics (IRSI), Proceedings of the CHI 2016-Workshop: Crowd dynamics: Exploring conflicts and contradictions in crowdsourcing*, *13*(2), 9–14.

# TiCC PhD Series

1. Pashiera Barkhuysen. *Audiovisual Prosody in Interaction.* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 3 October 2008.
2. Ben Torben-Nielsen. *Dendritic Morphology: Function shapes Structure.* Promotores: H.J. van den Herik, E.O. Postma. Co-promotor: K.P. Tuyls. Tilburg, 3 December 2008.
3. Hans Stol. *A Framework for Evidence-based Policy making using IT.* Promotor: H.J. van den Herik. Tilburg, 21 January 2009.
4. Jeroen Geertzen. *Dialogue Act Recognition and Prediction: Explorations in Computational Dialogue Modelling.* Promotor: H. Bunt. Co-promotor: J.M.B. Terken. Tilburg, 11 February 2009.
5. Sander Canisius. *Structured Prediction for Natural Language Processing: A constrained Satisfaction Approach.* Promotores: A.P.J. van den Bosch, W. Daelemans. Tilburg, 13 February 2009.
6. Fritz Reul. *New Architectures in Computer Chess.* Promotor: H.J. van den Herik. Co-promotor: J.W.H.M. Uiterwijk. Tilburg, 17 June 2009.
7. Laurens van der Maaten. *Feature Extraction from Visual Data.* Promotores: E.O. Postma, H.J. van den Herik. Co-promotor: A.G. Lange. Tilburg, 23 June 2009 (cum laude).
8. Stephan Raaijmakers. *Multinomial Language Learning: Investigations into the Geometry of Language.* Promotores: W. Daelemans, A.P.J. van den Bosch. Tilburg, 1 December 2009.
9. Igor Berezhnoy. *Digital Analysis of Paintings.* Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 7 December 2009.
10. Toine Bogers. *Recommender Systems for Social Bookmarking.* Promotor: A.P.J. van den Bosch. Tilburg, 8 December 2009.
11. Sander Bakkes. *Rapid Adaptation of Video Game AI.* Promotor: H.J. van den Herik. Co-promotor: P. Spronck. Tilburg, 3 March 2010.
12. Maria Mos. *Complex Lexical Items.* Promotor: A.P.J. van den Bosch. Co-promotores: A. Vermeer, A. Backus. Tilburg, 12 May 2010 (in collaboration with the Department of Language and Culture Studies).

13. Marieke van Erp. *Accessing Natural History: Discoveries in Data Cleaning, Structuring, and Retrieval.* Promotor: A.P.J. van den Bosch. Co-promotor: P.K. Lendvai. Tilburg, 30 June 2010.

14. Edwin Commandeur. *Implicit Causality and Implicit Consequentiality in Language Comprehension.* Promotores: L.G.M. Noordman, W. Vonk. Co-promotor: R. Cozijn. Tilburg, 30 June 2010.

15. Bart Bogaert. *Cloud Content Contention.* Promotores: H.J. van den Herik, E.O. Postma. Tilburg, 30 March 2011.

16. Xiaoyu Mao. *Airport under Control: Multiagent Scheduling for Airport Ground Handling.* Promotores: H.J. van den Herik, E.O. Postma. Co-promotores: N. Roos, A. Salden. Tilburg, 25 May 2011.

17. Olga Petukhova. *Multidimensional Dialogue Modelling.* Promotor: H. Bunt. Tilburg, 1 September 2011.

18. Lisette Mol. *Language in the Hands.* Promotores: E.J. Krahmer, A.A. Maes, M.G.J. Swerts. Tilburg, 7 November 2011 (cum laude).

19. Herman Stehouwer. *Statistical Language Models for Alternative Sequence Selection.* Promotores: A.P.J. van den Bosch, H.J. van den Herik. Co-promotor: M.M. van Zaanen. Tilburg, 7 December 2011.

20. Terry Kakeeto-Aelen. *Relationship Marketing for SMEs in Uganda.* Promotores: J. Chr. van Dalen, H.J. van den Herik. Co-promotor: B.A. van de Walle. Tilburg, 1 February 2012.

21. Suleman Shahid. *Fun & Face: Exploring Non-verbal Expressions of Emotion During Playful Interactions.* Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 25 May 2012.

22. Thijs Vis. *Intelligence, Politie en Veiligheidsdienst: Verenigbare grootheden?* Promotores: T.A. de Roos, H.J. van den Herik, A.C.M. Spapens. Tilburg, 6 June 2012 (in collaboration with the Tilburg School of Law).

23. Nancy Pascall. *Engendering Technology Empowering Women.* Promotores: H.J. van den Herik, M. Diocaretz. Tilburg, 19 November 2012.

24. Agus Gunawan. *Information Access for SMEs in Indonesia.* Promotor: H.J. van den Herik. Co-promotores: M. Wahdan, B.A. Van de Walle. Tilburg, 19 December 2012.

25. Giel van Lankveld. *Quantifying Individual Player Differences.* Promotores: H.J. van den Herik, A.R. Arntz. Co-promotor: P. Spronck. Tilburg, 27 February 2013.

26. Sander Wubben. *Text-to-text Generation by Monolingual Machine Translation.* Promotores: E.J. Krahmer, A.P.J. van den Bosch, H. Bunt. Tilburg, 5 June 2013.

27. Jeroen Janssens. *Outlier Selection and One-class Classification.* Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 11 June 2013.

28. Martijn Balsters. *Expression and Perception of Emotions: The Case of Depression, Sadness and Fear.* Promotores: E.J. Krahmer, M.G.J. Swerts, A.J.J.M. Vingerhoets. Tilburg, 25 June 2013.

29. Lisanne van Weelden. *Metaphor in Good Shape.* Promotor: A.A. Maes. Co-promotor: J. Schilperoord. Tilburg, 28 June 2013.

30. Ruud Koolen. *Need I say more? On Overspecification in Definite Reference.* Promotores: E.J. Krahmer and M.G.J. Swerts. Tilburg, 20 September 2013.

31. J. Douglas Mastin. *Exploring Infant Engagement, Language Socialization and Vocabulary Development: A Study of Rural and Urban Communities in Mozambique.* Promotor: A.A. Maes. Co-promotor: P.A. Vogt. Tilburg, 11 October 2013.

32. Philip C. Jackson. Jr. *Toward Human-level Artificial Intelligence — Representation and Computation of Meaning in Natural Language.* Promotores: H.C. Bunt, W.P.M. Daelemans. Tilburg, 22 April 2014.

33. Jorrig Vogels. *Referential Choices in Language Production: The Role of Accessibility.* Promotores: A.A. Maes, E.J. Krahmer. Tilburg, 23 April 2014 (cum laude).

34. Peter de Kock. *Anticipating Criminal Behaviour.* Promotores: H.J. van den Herik, J.C. Scholtes. Co-promotor: P. Spronck. Tilburg, 10 September 2014.

35. Constantijn Kaland. *Prosodic Marking of Semantic Contrasts: Do Speakers Adapt to Addressees?* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 1 October 2014.

36. Jasmina Marić. *Web Communities, Immigration and Social Capital.* Promotor: H.J. van den Herik. Co-promotores: R. Cozijn, M. Spotti. Tilburg, 18 November 2014.

37. Pauline Meesters. *Intelligent Blauw.* Promotores: H.J. van den Herik, T.A. de Roos. Tilburg, 1 December 2014.

38. Mandy Visser. *Better use your Head: How People learn to signal Emotions in Social Contexts.* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 10 June 2015.

39. Sterling Hutchinson. *How Symbolic and Embodied Representations Work in Concert.* Promotores: M.M. Louwerse, E.O. Postma. Tilburg, 30 June 2015.

40. Marieke Hoetjes. *Talking hands: Reference in speech, gesture and sign.* Promotores: E.J. Krahmer and M.G.J. Swerts. Tilburg, 7 October 2015.

41. Elisabeth Lubinga. *Stop HIV/AIDS. Start Talking? The Effects of Rhetorical figures in Health Messages on Interpersonal Discussions among South African Adolescents.* Promotores: A.A. Maes, C.J.M. Jansen. Tilburg, 16 October 2015.

42. Janet Bagorogoza. *Knowledge Management and High Performance: The Uganda Financial Institutions Models for HPO.* Promotores: H.J. van den Herik. Co-promotores: A.A. de Waal, B.A. Van de Walle. Tilburg, 24 November 2015.

43. Hans Westerbeek. *Visual Realism: Exploring Effects on Memory, Language Production, Comprehension, and Preference.* Promotores: A.A. Maes, M.G.J. Swerts. Co-promotor: M.A.A. van Amelsvoort. Tilburg, 10 February 2016.

44. Matje van de Camp. *A Link to the Past: Constructing Historical Social Networks from Unstructured Data.* Promotores: A.P.J. van den Bosch, E.O. Postma. Tilburg, 2 March 2016.

45. Annemarie Quispel. *Data for all: How Professionals and Non-professionals in Design use and evaluate Information Visualizations.* Promotor: A.A. Maes. Co-promotor: J. Schilperoord. Tilburg, 15 June 2016.

46. Rick Tillman. *Language Matters: The Influence of Language and Language use on Cognition.* Promotores: M.M. Louwerse, E.O. Postma. Tilburg, 30 June 2016.

47. Ruud Mattheij. *The Eyes have it.* Promoteres: E.O. Postma, H. J. Van den Herik, P.H.M. Spronck. Tilburg, 5 October 2016.

48. Marten Pijl. *Tracking of Human Motion over Time.* Promotores: E. H. L. Aarts, M. M. Louwerse. Co-promotor: J. H. M. Korst. Tilburg, 14 December 2016.

49. Yevgen Matusevych. *Learning Constructions from Bilingual Exposure: Computational Studies of Argument Structure Acquisition.* Promotor: A.M. Backus. Co-promotor: A. Alishahi. Tilburg, 19 December 2016.

50. Karin van Nispen. *What can People with Aphasia communicate with their Hands? A Study of Representation Techniques in Pantomime and Co-speech Gesture.* Promotor: E.J. Krahmer. Co-promotor: M. van de Sandt-Koenderman. Tilburg, 19 December 2016.

51. Adriana Baltaretu. *Speaking of Landmarks: How Visual Information influences Reference in Spatial Domains.* Promotores: A.A. Maes, E.J. Krahmer. Tilburg, 22 December 2016.

52. Mohamed Abbadi. *Casanova 2: A Domain Specific Language for General Game Development.* Promotores: A.A. Maes, P.H.M. Spronck, A. Cortesi. Co-promotor: G. Maggiore. Tilburg, 10 March 2017.

53. Shoshannah Tekofsky. *You are Who you Play you are: Modelling Player traits from Video Game Behavior.* Promotores: E.O. Postma, P.H.M. Spronck. Tilburg, 19 June 2017.

54. Adel Alhuraibi. *From IT-business Strategic Alignment to Performance: A moderated Mediation Model of Social Innovation, and Enterprise Governance of IT.* Promotores: H.J. van den Herik, B.A. Van de Walle. Co-promotor: S. Ankolekar. Tilburg, 26 September 2017.

55. Wilma Latuny. *The Power of Facial Expressions.* Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 29 September 2017.

56. Sylvia Huwaë. *Different Cultures, different Selves? Suppression of Emotions and Reactions to Transgressions across Cultures.* Promotores: E.J. Krahmer, J. Schaafsma. Tilburg, 11 October, 2017.

57. Mariana Serras Pereira. *A Multimodal Approach to Children's deceptive Behavior.* Promotor: M.G.J. Swerts. Co-promotor: S. Shahid. Tilburg, 10 January, 2018.

58. Emmelyn Croes. *Meeting Face-to-Face online: The Effects of Video-mediated Communication on Relationship Formation.* Promotores: E.J. Krahmer, M. Antheunis. Co-promotor A.P. Schouten. Tilburg, 28 March 2018.

59. Lieke van Maastricht. *Second Language Prosody: Intonation and Rhythm in Production and Perception.* Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 9 May 2018.

60. Nanne van Noord. *Learning Visual Representations of Style.* Promotores: E.O. Postma, M. Louwerse. Tilburg, 16 May 2018.

61. Ingrid Masson Carro. *Handmade: On the Cognitive origins of Gestural Representations.* Promotor: E.J. Krahmer. Co-promotor: M.B. Goudbeek. Tilburg, 25 June 2018.

62. Bart Joosten. *Detecting Social Signals with Spatiotemporal Gabor Filters.* Promotores: E.J. Krahmer, E.O. Postma. Tilburg, 29 June 2018.

63. Yan Gu. *Chinese Hands of Time: The Effects of Language and Culture on Temporal Gestures and Spatio-temporal Reasoning.* Promotor: M.G.J. Swerts. Co-promotores: M.W. Hoetjes, R. Cozijn. Tilburg, 5 June 2018.

64. Thiago Castro Ferreira. *Advances in Natural Language Generation: Generating varied Outputs from Semantic Inputs.* Promotor: E.J. Krahmer. Co-promotor: S. Wubben. Tilburg, 19 September 2018.

65. Yu Gu. *Automatic Emotion Recognition from Mandarin Speech.* Promotores: E.O. Postma, H.J. van den Herik, H.X. Lin. Tilburg, 28 November 2018.

66. Francesco Di Giacomo. *Metacasanova: A High-performance Meta-compiler for Domain-specific Languages.* Promotores: P.H.M. Spronck, A. Cortesi, E.O. Postma, Tilburg, 19 November 2018.

67. Ákos Kádár. *Learning Visually grounded and Multilingual Representations.* Promotores: E.O. Postma, A. Alishahi. Co-promotor: G.A. Chrupala. Tilburg, 13 November 2019.

68. Phoebe Mui. *The Many Faces of Smiling: Social and Cultural factors in the Display and Perception of Smiles.* Promotor: M.G.J. Swerts. Co-promotor: M.B. Goudbeek. Tilburg, 18 December 2019.

69. Véronique Verhagen. *Illuminating Variation: Individual Differences in Entrenchment of Multi-word Units.* Promotor: A.M. Backus. Co-promotores: M.B.J. Mos, J. Schilperoord. Tilburg, 10 January 2020 (cum laude).

70. Debby Damen. *Taking Perspective in Communication: Exploring what it takes to change Perspectives.* Promotor: E.J. Krahmer. Co-promotores: M.A.A. van Amelvoort, P.J. van der Wijst. Tilburg, 4 November 2020.

71. Alain Hong. *Women in the lead: Gender, Leadership Emergence, and Negotiation Behavior from a Social Role Perspective.* Promotor: J. Schaafsma. Co-promotor: P.J. van der Wijst. Tilburg, 3 June 2020.

72. Chrissy Cook. *Online gaming and trolling.* Promotores: J. Schaafsma, M. Antheunis. Tilburg, 22 January 2021.

73. Nadine Braun. *Affective Words and the Company They Keep: Investigating the interplay of emotion and language.* Promotor: E.J. Krahmer. Co-promotor: M.B. Goudbeek. Tilburg, 29 March 2021.

74. Yueqiao Han. *Chinese Tones: Can you Listen with your Eyes? The Influence of Visual Information on Auditory Perception of Chinese Tones.* Promotor: M.G.J. Swerts. Co-promotores: M.B.J. Mos, M.B. Goudbeek. Tilburg, 18 June, 2021.

75. Tess van der Zanden. *Language Use and Impression Formation: The Effects of Linguistic Cues in Online Dating Profiles.* Promotor: E.J. Krahmer. Co-promotores: M.B.J. Mos, A.P. Schouten. Tilburg, 22 October 2021.

76. Janneke van der Loo. *Mastering the Art of Academic Writing: Comparing the Effectiveness of Observational Learning and Learning by Doing.* Promotor: E.J. Krahmer. Co-promotor: M.A.A. van Amelsvoort. Tilburg, 1 December 2021.

77. Charlotte Out. *Does Emotion shape Language? Studies on the Influence of Affective State on Interactive Language Production.* Promotor: E.J. Krahmer. Co-promotor: M.B. Goudbeek. Tilburg, 16 December, 2021.

78. Jan de Wit. *Robots that Gesture, and their Potential as Second Language Tutors for Children.* Promotor: E.J. Krahmer. Co-promotor: P.A. Vogt. Tilburg, 28 January, 2022.