universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

Ănter- and Intra-speaker Variation in Multi-modal Task Descriptions and Implications for Human-Robot Interactionñ

verfasst von / submitted by

Mag. Stephanie Gross, MSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktorin der Philosophie (Dr. phil.)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 792 327
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Sprachwissenschaft
Betreut von / Supervisors:	Univ.-Prof. Dr. Daniel Büring Univ.-Prof. Matthias Scheutz PhD, PhD

Contents

Preface

During my studies of Linguistics and Cognitive Sciences, I was highly interested in the kind of linguistic knowledge, cognitive mechanisms, sensory and action capabilities artificial systems need as basic prerequisite in order to be able to communicate with humans. Next to my studies, I started an internship at the Austrian Research Institute for Artificial Intelligence (OFAI) in 2009, which gave me the opportunity to get acquainted with and work in different areas in the field of computational linguistics and cognitive sciences. In the very supporting working environment at OFAI and during inspiring discussions with my colleagues, my interest in the interplay of verbal and non-verbal situated communication grew.

Nowadays, employing commercial speech recognition software is still very cumbersome. Solely linguistic knowledge is in most circumstances not sufficient for artificial systems to communicate in ways natural for their human interlocutors. To successfully interact with humans, artificial systems have to share representations about events, actions, objects, agents etc. with their communication partner. On the other hand, they have to be able to deal with a broad variation of verbal and non-verbal communication cues. Depending on the communication context, different processing capabilities are needed by the interlocutors, e.g., when talking on the phone, talking face to face about imaginative events, or situated task descriptions. Scenarios where robots are more and more used in our everyday-lifes are household robotic assistants, museum robots, companions or assistive robots for the elderly or people with disabilities, etc. In these contexts, robots need to interact with and learn from non-expert humans in order to be able to deal with changing contexts. Situated task descriptions are a form of communication, which might frequently occur.

In situated task descriptions, dealing with the multi-modality of the interlocutor might be as or even more relevant than sharing representations with the communication partner. As the objects and actions relevant for the task are in the shared environment of the interlocutors, it might be more important to be able to resolve a verbal referring expression, e.g., "the tube", to an entity that might be called a tube

in the visual field than to share a similar cognitive representation of "tube". Thus, I am interested in the interplay of verbal and non-verbal communication. My work is driven by the conjecture that language might not play a primary, but a secondary role in situated task descriptions.

The viewpoint taken in the research questions is what kind of multi-modality an artificial system, e.g., a robot, would have to deal with, if it were in the learner's position. Thus, its main objectives are to (i) collect empirical human situated, task-based interaction, (ii) analyse the multi-modal data for processes, that might be challenging for robots and their variation between and within task descriptions, with a focus on references to objects, (iii) extract general principles on the interplay of the different modalities, and (iv) formulate design suggestions for robot architectures on how to deal with the accordant principles.

As a first step, I analysed data which researchers at the Technical University Munich (TUM) at the Institute for Information-Oriented Control (ITR) collected and provided. In this data, 19 people working at ITR explained a task to a future instructor. The setting was borrowed from an actual robot-learning task and I used the data for a pilot study to (i) gain first insights in the multi-modality of human-human task descriptions and variations between and within task descriptions, (ii) employ a framework of embodied language comprehension on the data and investigate whether it is suitable for situated task descriptions, and, based on the first two aspects, (iii) motivate a setup of a data collection, including a larger number of participants, tasks and technical recording equipment.

The quantitative and qualitative analysis of the data shows that multi-modal communication plays a crucial role in situated task descriptions. If only the verbal part of task descriptions is used for interpretation, important information for successfully conducting the task is missing. Information transmitted via language (i) was often erroneous, disfluent, and vague, (ii) variations in wording between and within speakers occurred, ranging from specific to unspecific words, (iii) the perspective taken by the teacher, e.g., "I'", "you", varied between and within teachers, and (iv) spatial indexicals frequently occurred, e.g., 'here', 'like this'. However, in parallel to these utterances, eye gaze and gestures transmitted crucial information for resolving references. These results underpin the importance of non-verbal communication cues in human task descriptions. Thus, linguistic and visual information – especially gaze and gestures – need to be incrementally incorporated in a robot architecture to resolve referents of unspecific noun phrases or pronouns lacking verbal antecedents. The results of the pilot study are published in Schreitter & Krenn

(2013b)[1]. In Schreitter & Krenn (2013a), we discussed the suitability of employing a framework of embodied language comprehension on the empirical data.

Based on the results of the pilot study, I developed a setup including four different tasks with four different foci: the first task includes voiced object names to allow for an analysis of information structure and there is no learner present. The second is a collaborative task with one very salient object. In the third task description, the learner is only observing and listening and it includes a variety of different, although similar objects. The fourth is a navigation task, where the learner has to follow instructions and navigate to a certain location.

In April 2013, I spent four weeks at ITR to collect the data. The reason for collecting the data at TUM was that at ITR, I was able to not only collect human-human but also human-robot data, by employing a Wizard-of-Oz setting. Matthias Rambow at ITR supported me greatly with the technical aspects of the WoZ-Setting as well as with the recording of motion and force data during the data collection. We collected data of 22 human teachers explaining four different tasks to either a human or a robot learner. The tasks were designed to investigate on which channels, such as language, eye gaze or gesture, relevant information is transmitted during a task description. Thus, all in all the corpus comprises 88 German recordings. In 22 recordings the descriptions are directed towards the camera, in 54 recordings the task descriptions are directed towards a human learner, and in 12 they are directed towards a robot learner. A description of the data collection was published in Gross & Krenn (2016).

After converting and annotating the data, I first conducted an explorative analysis similar to the one of the pilot study, to investigate which aspects of multi-modal task descriptions are especially challenging for human-robot interaction, and which aspects might allow for an automatic analysis and can be used to ease information processing. Especially the variation in verbal referring expressions was surprisingly high. A part of these results was published in Schreitter & Krenn (2014).

Due to the high variation of verbal references, I took a closer look at referring expressions to objects and their multi-modality, driven by the following research questions: how high is the inter- and intra-speaker variation when referring to one individual object, how often are references to objects underspecified, and what is the role of eye gaze and gestures? A summary of the results was published in Gross et al. (2016).

The third part of the extensive data analysis was to investigate the verbal part of referring expressions based on their linguistic form. The goal of this analysis was to

[1]Last name changed to Gross from Schreitter in 2016.

identify non-verbal cues, which might be very relevant for a certain linguistic form, e.g., pronouns, while for other linguistic forms, e.g., noun phrases, other non-verbal cues are of relevance.

Based on these results, I then extracted challenges and provided suggestions for developing a multi-modal reference-resolution mechanism for robots in a shared environment with a human interlocutor.

The ideas, realisation, and analysis of the work mentioned above stem from myself, however, I received valuable input from my supervisors Brigitte Krenn and Matthias Scheutz.

Some of the results have already been considered in the development of an open world reference resolution algorithm and implemented in a robot architecture. At the beginning of 2015, I had the opportunity to spend two months at the Human Robot Interaction Laboratory at Tufts University, deepen my knowledge about the robot architecture DIARC, collect English human-human interaction data and collaborate with researchers at the institute. A result of interesting and productive conversations is the adaptation of the Givenness Hierarchy by Gundel et al. (1993). Tom Williams, Saurav Acharya, and Matthias Scheutz were working on implementing a version of the Givenness Hierarchy and we adapted the model for reference resolution to the results of my analysis in order to enable it to also deal with situated task descriptions. Subsequent to their implementation of the algorithm, we evaluated the model on a subset of English translations of the data I collected. For the evaluation, I provided data material and general input. The results are published in Williams et al. (2015) and Williams et al. (2016).

This collaborate work on the adapted version of the Givenness Hierarchy, its implementation and the evaluation on situated task descriptions is very much in line with my research in this thesis. It includes results from multi-modal situated human task descriptions with the aim to enhance reference resolution in human-robot interaction.

At this point, I want to acknowledge the support from various sides which made this thesis possible. In particular, I am very grateful to my supervisor Matthias Scheutz for his guidance, helpful comments and encouragement throughout my research for this work despite the geographical distance. No less important, I would like to thank my supervisor Daniel Büring for his support, input and sympathetic ear.

During this work, I was funded by the DOC Fellowship of the Austrian Academy of Sciences and employed at OFAI. The fellowship as well as the fruitful working environment at OFAI gave me the freedom to pursue my research interests and take

new initiatives. In particular Brigitte Krenn, my team-lead in the Language and Interaction Technologies Group, supported and encouraged me with her enthusiasm and helpful suggestions and comments before and throughout my work on this thesis.

I would also like to thank the ITR at TUM, especially Sandra Hirche and Matthias Rambow, for their support with the WoZ-Setup and recording the data and Katharina Kranawetter for annotating parts of the data collection. And finally, I want to thank all members of the Human Robot Interaction Laboratory at Tufts University, especially Tom Williams, for inspiring discussions.

Vienna, October 2016

Chapter 1

Introduction

In situated task descriptions, human interlocutors are embedded in a physical world and refer to objects and actions using both linguistic and non-linguistic forms of communication. The linguistic information often does not comprise all information necessary for understanding a task. When robots have to learn tasks from humans in the future, we need to better understand the various relevant aspects of human multi-modal task descriptions and how verbal and visual information can be detected and integrated.

In the following, the motivation for this thesis as well as research aims and objectives will be outlined, followed by a presentation of the methodology used in this work. The chapter will conclude with a summary of the thesis structure.

1.1 Motivation

Imagine a robot that can analyse, interpret, and learn from task-oriented presentations where a human teacher shows a task to the robot learner and explains what she/he is doing by means of task-accompanying speech. For robots to be able to deal with the multi-modal complexity of human communication, we need to better understand general principles of human task-based descriptions within a shared environment in order to distil the critical interaction principles that have to be integrated into robotic control architectures (i.e., the software and hardware framework for controlling a robot).

Application domains for human-robot interaction. When designing robot architectures, it is not possible for the expert designer to foresee, what the communication and application contexts of the robot will exactly look like. Thus, it is important for a robot to be able to adapt to new contexts. In this case, a non-expert

human will need to take the role of the instructor in order to teach the robot how to handle a new context. In order to be able to learn from human instructors, it is essential to equip robots with natural language capabilities. For successful interaction, artificial systems need to share representations about events, actions, objects, agents etc. with their human interlocutor. In case they are co-located with their communication partner and the interaction is task-based, they also have to be able to deal with a broad variation of verbal and non-verbal communication cues. Depending on the communication context, different processing capabilities are needed by the robot.

Actual application domains, in which multi-modal human-robot interaction is needed, include household robotic assistants (e.g., Ciocarlie et al., 2014), assistive robotics and companions for special groups of people, such as older adults (e.g., Fischinger et al., 2016), persuasive robotics (e.g., Ham et al., 2015), robotic educational assistants (e.g., Fridin, 2014), museum robots (e.g., Yamazaki et al., 2012), tour guides (e.g., Karreman et al., 2013), robotic wheelchairs (e.g., Tellex & Roy, 2006), and companion robots (e.g., Dautenhahn et al., 2006).

Task-based descriptions within a shared environment are frequent communication contexts in all of these application scenarios, e.g., when a human is instructing a robot how to conduct a new task. In situated task descriptions, identifying and interpreting multi-modal cues by the interlocutor might be as or even more relevant than sharing representations with the communication partner for, e.g., the resolution of referring expressions to objects. If the objects are relevant for the task, they are also co-present and it is more important to resolve the reference, might it be "thing" or "tube", to an object in the visual field of the robot than to share a representation of "thing" or "tube" with the interlocutor. Thus, this thesis focuses on the interplay of verbal and non-verbal communication in order to grasp the information necessary for the task, with a focus on references to objects.

Human situated task descriptions. Human instructors use not only speech, but various multi-modal communication cues such as eye gaze and gestures, when showing and explaining a task to a learner, especially when the learner is physically co-present (see McNeill, 1992; Kendon, 2004; H. H. Clark & Krych, 2004; Hanna & Brennan, 2007). While language could theoretically be used as the major, possibly even only information channel, it will often be underspecified and is heterogeneously used by individual speakers (see Furnas et al., 1984, 1987; Brennan, 1996).

A task description such as the following (taken from the data collected for this thesis) emphasizes the importance of non-verbal cues, especially gesture in this case.

The description begins the following way:

> "Yes, now we have this thing here and our task is that we turn it around, which means that the part at your side is then at my side and the one at my side is then over there."
> *Ja, also wir haben hier dieses Ding und unsere Aufgabe ist, dass wir das einmal herumdrehen, das heißt, dass die Seite bei dir bei mir ist und die bei mir da drüben ist.*

The object manipulated in this task is a board with two handles. This introduction to the task is accompanied by six gestures. The first one is a pointing gesture to the board while uttering "the thing", the second one is a iconic gesture depicting the action and direction of turning the board. During the last four gestures the instructor is pointing at the location where the accordant handle of the board is located at the time and where it will be after the turning-action. Thus, these six references mentioned by the participant can only be resolved via gestures.

In task descriptions, language might even take a subordinate, guiding role when task-relevant objects and actions do not have to be inferred from natural language expressions, but can be directly observed. In that case, non-verbal cues such as gestures and gaze of the speaker are often employed as indicative acts during communication (H. H. Clark, 2003; Brennan, 2000), even though humans could communicate the intended information through language alone. Incorporating visual information is thus a necessary prerequisite to deal with situated task descriptions.

In order to investigate human task descriptions in more detail, an experimental setup was designed, and data was collected and analysed where a teacher explains and shows different tasks to a learner. By letting different people explain the same tasks, insights can be gained about how humans naturally structure and present information and the variation between and within task descriptions. Thus, the results are an important basis for what a robot would have to deal with if it were in the learner's position.

Limitations of computational models. There exists converging psycholinguistic evidence that pointing, eye gaze, placing objects etc. play an important role during language understanding in humans (H. H. Clark, 2003; H. H. Clark & Krych, 2004; Brennan, 2000). Computational models aimed at understanding human language need to account for its multi-modal complexity. Despite the wealth of empirical research on referring expressions in psycholinguistics and the work on developing models in computational linguistics, these two fields proceed with little mutual influence (see Van Deemter et al. (2012) and Gatt et al. (2014)).

However, in most approaches, the interaction model is handled separately from the discourse model. Most computational approaches for resolving and generating referring expressions focus only on language and leave non-verbal communicative cues aside, e.g., the Centering Theory by Grosz et al. (1995) or the Incremental Algorithm Dale & Reiter (1995) and some of their more recent adaptations (e.g., Krahmer & Theune (2002); Goudbeek & Krahmer (2012).

Some computational approaches, which take also non-verbal cues into account for the resolution of referring expressions, are for example Kehler (2000); Chai et al. (2006); Kranstedt et al. (2006); Van der Sluis & Krahmer (2007); Prasov & Chai (2008); Lemaignan et al. (2012); Admoni et al. (2014); Huang & Mutlu (2014).

Non-verbal cues in human communication need to be tightly integrated with language. Hence, both verbal and non-verbal processing need to be handled flexibly and might contribute essential, even if not the entire information for reference resolution. Only the integrated channels allow for the resolution of references. Non-verbal cues accounted for in current models of reference resolution include up to three different cues: objects in the visual field, eye gaze, and gestures. None of the above mentioned models propose solutions for how to deal with verbal and non-verbal aspects of inherently multi-modal situated communication, i.e., which non-verbal cues need to be accounted for, as well as their reliability and interlinkage for automatic reference resolution. Hence, it is critical that we develop more comprehensive computational models of human reference resolution in task-based contexts where instructor and instructee are co-located. This will not only inform the theory of situated natural language interactions, but also provide important design suggestions and constraints for the development of artificial agents that interact with humans in such contexts.

And although the objects in the visual field, eye gaze, and gestures of the interlocutor are transmitting crucial information to resolve references in situated task descriptions, the approach presented in this thesis includes a more extensive and explorative analysis, including additional cues. First, general principles of human situated task descriptions are extracted, in order to identify general challenges for human-robot interaction. Second, with a focus on the resolution of referring expressions to objects, in particular (i) relevant verbal and non-verbal cues beyond the ones mentioned above were extracted, (ii) the reliability of the different modalities was investigated, as well as (iii) the variation of verbal referring expressions, and (iv) the interplay of linguistic forms and specific non-verbal cues.

1.2 Research Aim and Objectives

The major goal of this thesis is to identify what kind of variation and multi-modality a robot would have to deal with when a human instructor explains a task within a shared environment, and formulate design principles for robotic systems in order to deal with this variation and multi-modality. Substantial non-verbal modalities, variations and congruences between and within instructors need to be identified, with regard to how task-relevant information is transmitted. The presented work is also driven by the conjecture that in situated referential interaction, human instructors vary vastly in how they structure and present a task to a learner, and that language might not play a primary, but a secondary role in situated task descriptions.

The main objectives of this thesis are:

- to design a setup for an empirical human-human and human-robot data collection and collect data, suitable to investigate situated task interaction;
- to analyse the multi-modal data for processes, that might be challenging for robots and their variation between and within task descriptions;
- to investigate the applicability of a general cognitive model for human embodied language comprehension on situated task descriptions;
- to extract general principles on the interplay of different modalities, with focus on reference resolution to objects;
- to compare the extracted principles with an already existing model for reference resolution;
- to formulate design suggestions for robot architectures on how to deal with the accordant principles.

Preceding to designing a setup and collecting data for the thesis, a smaller data set collected by researchers at the Institute for Information-Oriented Control (ITR) at the Technical University Munich was analysed. The study was used to (i) gain first insights in the multi-modality of human-human task descriptions, (ii) investigate whether a theoretical framework of embodied language comprehension (the "Immersed Experiencer Framework" by (Zwaan, 2004)) is suitable for situated task descriptions, and (iii) motivate a setup of a data collection, including a larger number of participants, tasks, and technical recording equipment.

The design of the setup is chosen in a way to allow for a comparison of how individual instructors vary or concur in how they structure information during a task description. The data collection comprises of four tasks with different foci. The first task is directed towards a camera and it includes voiced object names to allow for an analysis of information structure. The second task is collaboratively

conducted and contains one salient object. In the third task, the learner is observing, while the instructor is conducting and explaining the task. It includes a variety of different, similar objects. In the fourth task, the learner has to follow instructions and navigate to a certain location.

The analysis of the data is threefold. The first explorative part is guided by the following research questions: Through which channels, such as language, eye gaze or gesture, is relevant information transmitted? What is the variation in conveying respective information between instructors and tasks, but also within a task description, i.e., the inter- and intra-speaker variation? What are differences in how a task is transmitted between human-human and human-robot dyads? In addition to an explorative analysis, multi-modal reference resolution to objects is investigated: (i) the variation in the choice of nouns denoting one specific object, (ii) lexical underspecification for conceptual content, and (iii) the role of eye gaze and gestures when uttering referring expressions. The third part of the data analysis is dedicated to the connection of the linguistic form and non-verbal cues. In addition to eye gaze and gesture, other non-verbal cues are investigated which are needed to allow for a comprehensive resolution of all referring expressions to objects. Based on the results, general principles of multi-modal task descriptions are extracted.

The main goal of the thesis is to formulate design suggestions based on the extracted results, e.g., how to deal with the variation of expressions referring to one specific object, underspecified object references, or the multi-modality of referring expressions.

Part of the extracted principles were used to develop an adapted version of the Givenness Hierarchy by Gundel et al. (1993) (a model for reference resolution) to situated task descriptions. This work was conducted in collaboration with Tom Williams, Saurav Acharya, and Matthias Scheutz.

The work presented in this thesis differs from previous investigations in the depth of the analysis of multi-modal referring expressions in situated human-human task descriptions and the non-verbal cues accounted for in reference resolution. Also the reliability of different cues for resolving references to objects as well as the order in which they need to be processed according to their linguistic form will be integrated in the resulting design suggestions for robot architectures.

1.3 Methodology

The experimental setup focuses on human-human dyads to inform human-robot interaction. Although human-robot interaction is recorded as well, the number of

recordings is rather small. I am aware that the results of human-human interaction can not be transmitted one-to-one to human-robot interaction. There is a growing body of literature on the influence of the morphology of robots on the users' behaviours (see Vollmer et al., 2009; Pitsch et al., 2012). However, humans naturally employ a wide range of variation in verbal and non-verbal referring behaviour in inherently multi-modal situated communication (see Brennan, 1996; H. H. Clark & Krych, 2004; Hanna & Tanenhaus, 2004). Using human-human interactions is driven by the assumption that humans expect at the least the sophisticated communicative behaviour from humans as they do from robot interlocutors. By considering human-human interaction, the whole spectrum of human referring behaviour is embraced for robot architecture design, independent of the morphology of the robot.

For the analysis of the corpus, a combination of qualitative and quantitative methods is used. In order to identify relevant channels through which information is transmitted, as well as their interplay, an inductive approach is employed, in combination with frequencies of occurrences. In addition to a pilot study, three analyses of empirical data are presented. As the thesis progresses, the research questions will be narrowed down and focused more and more on sub-problems based on the results of the preceding analyses. However, results which will not be part of the subsequent analyses of the data are still valuable and will be also accounted for in the design suggestions for robot architectures.

Python and R were used for an automatic analysis of the data.

1.4 Thesis Structure

Chapter 2, **Theoretical and Empirical Background**, reviews and discusses relevant research literature. It covers aspects of human-human communication, which are potential challenges for automatic speech processing and thus also for human-robot interaction.

In Chapter 3, **Multi-modal Human-Robot Interaction**, general challenges for robots in situated human-robot interaction are discussed, as a first step towards design principles for multi-modal human-robot communication. Current computational approaches to multi-modal reference resolution are reviewed and discussed. The chapter closes with a discussion of data collections, suitable for research in human-robot interaction, as well as annotation schemes and tools.

In Chapter 4, the **Pilot Study** is introduced, including a discussion of the suitability of employing a general model of embodied language comprehension on the data. First results on multi-modal human-human task descriptions are presented, as well as an annotation scheme including these first results plus parts of the model of embodied language comprehension. The chapter will conclude with the resulting motivation for developing a more comprehensive setup for data collection.

Chapter 5, **Empirical Study on Human Multi-modal Task Descriptions** is concerned with the empirical data collection and analysis, including the annotation scheme, technical tools used for analysis and research interests. The analysis of the data is structured in three parts. First, inter- and intra-speaker variation is exploratively investigated. Second, reference resolution to objects is analysed with a focus on the role of language, eye gaze, and gesture. Third, non-verbal cues are extracted in addition to eye gaze and gesture, and results are presented with regards to the interplay of linguistic forms and non-verbal modalities.

Having so far covered the empirical basis, Chapter 6, **Challenges and Architectural Design Suggestions for Robots in Multi-modal Human-Robot Interaction**, deals with requirements for a computational model for reference resolution to objects. It includes challenges such as the variation of expressions referring to one specific object, underspecified verbal referring expressions, and their multi-modality. Design suggestions are formulated, which depend on the accordant challenge.

Chapter 7, **Preliminary Implementation**, describes the development, validation and evaluation of an algorithm for situated open world reference resolution. This chapter summarizes collaborate work with Tom Williams, Saurav Acharya, and Matthias Scheutz from Tufts University. First, an algorithm is proposed, using the Givenness Hierarchy and adapted according to empirical results presented in Chapter 5. The algorithm is then evaluated on a subset of the data collection, also outlined in Chapter 5.

Finally, Chapter 6, **Conclusion**, summarizes and discusses the results and provides an outlook on future research perspectives.

Chapter 2

Theoretical and Empirical Background

In this work, the focus is on experimental data from human experiments to inform the development of mechanisms for robots to deal with multi-modal information transmitted by an instructor in situated task descriptions. In human-human task-based interaction, it is on the one hand necessary to share representations of objects, actions, and agents with the interlocutor in order to successfully communicate. Compelling evidence from embodied cognition has shown the importance of action and perception during language comprehension in humans. "Embodied" in this context refers to having a body and experiencing the world by means of it. On the other hand, people use a multitude of verbal and non-verbal behaviours such as communicative gestures, object manipulation gestures, gazes and nods accompanying their verbal utterances. Only the combination of the vocal and the gestural acts together provides the information necessary for the interlocutor to understand situated communication. Information transmitted via different channels, such as language, eye gaze or gesture, need to be identified, interpreted, and merged.

In this work, the focus is on situated task descriptions, as they are a form of communication, frequent in many different application fields of human-robot interaction. With regards to theoretical and empirical background, both of the above mentioned aspects are of importance; however, the focus is on the multi-modality of task descriptions within a shared environment. In this chapter, first, relevant literature on multi-modal aspects of human-human task-based interaction are discussed (Section 2.1), looking at general inter- and intra-speaker variation in language, verbal referring expressions, and multi-modal cues in situated interactions. Subsequently, important findings from embodied language comprehension are briefly summarized and the Immersed Experiencer Framework is introduced (Section 2.2). The chapter

will conclude with a summary and discussion in Section 2.3.

2.1 Situated Multi-modal Human-Human Interaction

During communication within a shared environment interlocutors refer to actual objects, agents, locations etc. In order to do so, Clark emphasizes that when communicating, people create different signs for their interlocutors H. H. Clark (1996). According to Peirce, there are three modalities of signs: icons, symbols, and indexes (Peirce & Buchler, 1955). They differ in the relation between the sign and the object. While a symbol is associated with its denoted object by a rule, an icon is linked to its referred object via a perceptual resemblance, and an index designates its referred object via an actual (e.g., spatial) connection. H. H. Clark (2003) calls this act of creating a sign *signalling* and defines three signalling methods:

Describing-as: Using symbols to signify categories of things
Demonstrating: Creating icons or selective depictions of things
Indicating: Forming indexes to individual things

If you consider a task description, all of these three signalling methods are likely to frequently occur. In an utterance such as "You need a hammer", "hammer" is produced as a symbol to signify a certain category of things. The action of hammering can be demonstrated by moving a hand up and down as if actually hammering. And in case there are two hammers in the visual field, one might utter "You need this hammer", while pointing at the referred object. H. H. Clark (2003) argues that placing objects "just in the right manner" is also a form of indicating. For example Person B instructs Person A in how assemble a piece of furniture. Person A needs to hammer and Person B brings one out of two hammers, e.g., the smaller one, and places it next to the workspace of Person A. By this act of placing, Person A is instructed which hammer to use.

Both icons and indexes inherently need visual cues in order to resolve the references. Using symbols such as in "You need a hammer" can be sufficient to resolve references, but for example in case there are two hammers, also additional information is needed. This multi-modal complexity needs to be accounted for when developing computational models aiming at understanding human language. In natural communication, the combination of verbal and non-verbal communication comprises the information necessary for understanding. During situated task descriptions, humans communicate by utterances, exhibiting, poising, pointing at,

placing, and orienting objects, and by eye gaze, head nods, and head shakes, all timed with precision, see (H. H. Clark & Krych, 2004).

2.1.1 General inter- and intra-speaker variation in language

In spontaneous, spoken language, there are several aspects where the interlocutors need to adapt to one another and need to include non-verbal cues for interpretation in order to grasp the meaning of what has been communicated. In the following, aspects of language are discussed, which are potential challenges for automatic speech processing and thus also for human-robot interaction.

Variation of content words. Humans often utter different words for one and the same object, action, etc. and given the average size of a human mental lexicon, the potential for variability in word choices is enormous. Furnas et al. (1987, 1984) named this phenomenon "the vocabulary problem". In their studies on human-computer dialogue, they found that two people producing the same term for the same command (e.g., delete a file on a computer) only ranged from 7-18%. In a study by Brennan (1996) on lexical variability in human-human dialogue, the likelihood that people choose the same terms for the same common objects (e.g., shoes, dogs, cars, fishes) as another instructor in another trial was only 10%. However, when two people repeatedly discussed the same object within a conversation, variability was relatively low. Reason for this was lexical entrainment, i.e., they came to use the same terms during the interaction. In human-robot interaction, the potential of different word choices by humans makes high demands on reference resolution.

Extending the study by Brennan (1996), in this thesis, lexical variation between and within situated task descriptions are investigated, as well as the amount of underspecified noun phrases. In contrast to the study by Brennan, there is no lexical entrainment in the data presented in this work, because there is mainly one person speaking while the other one is mainly listening.

Disfluencies effects. Characteristics of language, such as abandoned utterances, filler words or repairs frequently occur in natural speech production. Speakers become more disfluent when cognitive load increases. Bortfeld et al. (2001) investigated potential factors influencing fluency rates such as speakers' ages, task roles (instructor versus instructee), relationship between speakers and gender. The results showed that disfluency rates were higher when both speakers acted as instructors, and when complex domains were discussed, i.e., when cognitive load increased. Related to references to objects or actions, disfluencies increase for lexical retrievement

(i) when the word being planned is low in frequency (Beattie & Butterworth, 1979; Eisler, 1968), or (ii) when the object has not been mentioned recently or when it is unconventional and thus lexical retrievement is more difficult (Arnold & Tanenhaus, 2011).

Perspective taking. To interpret the perspective taken by the speaker can already be a challenge in human-human interaction. There has to be a permanent adaptation to the interlocutor in order to interpret utterances of an interlocutor, see Barr & Keysar (2006) for an overview. Despite the variation of content words (see Furnas et al., 1987, 1984; Brennan, 1996), there is also the need for perspective taking or the negotiation of meaning with regards to pronouns. Oshima-Takane et al. (1996) argues that children learn "I" and "you" not solely in child-caregiver interactions, but by observing others interacting with others. The authors provide for example evidence, that secondborn children were more advanced in pronoun production than firstborn children, while not differing in general language development.

Additionally, personal pronouns *I, we, you* can be used in many languages (amongst others German and English) as impersonal pronouns transmitting structural knowledge and general truths, see Kitagawa & Lehrer (1990) for an overview. In such sentences, the pronoun could be replaced by *one*, and in indirect speech the expected person shifts do not occur. These stylistic and rhetorical differences among impersonally used personal pronouns follow from their deictic use. This flexibility in the use of personal pronouns also poses a challenge for human-robot interaction.

Based on psycholinguistic studies, Brennan (2000) extracted a number of implications for both computational linguistics and human-computer interaction:

- Corpus data including systematic information about the task can be valuable for the development of dialogue systems.

- Language processing modules should not be based on the assumption that utterances are complete and well-formed.

- Non-propositional features of language should be included as well, such as timing or intonation.

- Computational dialogue systems should include resources to the negotiation of meaning, modelling context, recognizing which referring expressions are likely to index a certain entity. When a new referring expression is uttered, it could be marked as provisional before lexical entrainment might occur. By tracking

the already used forms of referring expressions in the discourse, agents can be enabled to use the same terms consistently to refer to the same object.

- Dialogue models should keep a structured record of jointly achieved contributions that is updated and revised incrementally.

Although these implications only focus on the linguistic part of referring expressions, they already highlight the need to deal with the above mentioned aspects of human-human interaction when developing natural language processing systems.

2.1.2 Verbal referring expressions

It is a central ability of human communication to be able to refer to and identify entities in a shared environment. Especially for initial referring expressions, language can be ambiguous and thus often requires coordination between interlocutors in order to be successful.

In most approaches, there is the general assumption that there is a direct relationship between the form of a referring expression and the accessibility of the referent in the addressee's discourse model (see Ariel, 2001; Gundel et al., 2012; Gatt et al., 2014). These referring expression can take different forms. (Gatt et al., 2014; Reiter et al., 2000) for example identify the following forms: (i) a full name (e.g., *Bill*), a pronoun (e.g., *it*) or a description (e.g., *the large blue aeroplane*), while in the Accessibility theory by Ariel, the accessibility scale contains 18 different markers ranging from a full name plus a modifier to a verbal person inflection (with a zero subject) and zero (all are higher accessibility markers than a full pronoun).

The Givenness Hierarchy by Gundel et al. (1993) spans six what they call "cognitive statuses" marked by different forms of referring expressions, see Table 2.1. Each level is contained by all lower levels, thus information that is in focus is also activated, familiar, etc.

Independent of the approach, the chosen form is influenced e.g., by whether the object is referred to for the first time (initial reference) within the discourse, or whether the object was mentioned before (subsequent reference). While full names are typically used for initial references, reduced forms, such as pronouns, are employed when the referent already has high salience (Reiter et al., 2000).

Verbal descriptions carry information which has to be identified and extracted by a listener, e.g., for pronoun resolution. Arnold et al. (2000) found evidence that gender and accessibility information influence referent consideration during the initial process of pronoun resolution. For resolving ambiguous pronouns, considerable

Table 2.1: Level, cognitive status and form in the Givenness Hierarchy

level	**cognitive status**	**form**
in focus	in focus of attention	{*it*}
activated	in working memory	{*that, this, this* N}
familiar	in long term memory	{*that* N}
uniquely identifiable	in long term memory or new	{*the* N}
referential	new	{indefinite *this* N}
type identifiable	new or hypothetical	{*a* N}

attention has been paid to heuristic strategies such as the first-mention account by Gernsbacher & Hargreaves (1988); Gernsbacher et al. (1989) and the subject-preference account by Crawley et al. (1990); Frederiksen (1981). In the first-mention account, the first mentioned noun phrase is the preferred antecedent of an ambiguous pronoun. The subject-preference account assumes that the preferred antecedent is the grammatical subject of the preceding clause or sentence. In an eye-tracking study Järvikivi et al. (2005) found evidence for both accounts.

Information transmitted within a verbal description might be not sufficient to resolve reference to a certain object, e.g., spatial indexicals without verbal antecedents, variations in wording, and omitted verbal references for objects, actions, and locations. In order to still be able to extract the content, information transmitted via the visual modality needs to be interlinked with information transmitted via the linguistic modality. Both gaze and gestures are important cues for establishing joint attention (Tomasello & Akhtar, 1995; H. H. Clark & Krych, 2004; Frischen et al., 2007).

Reference resolution or accessibility of potential referents is generally assumed to be related to focus of attention on certain entities in the discourse situation (see Almor, 1999; Dahan et al., 2002; Gundel et al., 2012). In general, discourse-old or given are considered more accessible than discourse-new entities (Chafe & Li, 1976; Prince, 1992). The traditional approach to investigate accessibility and reference resolution is via linguistic mentions. However, also visual presentation and inferring entities through association play an important role for referring to objects and actions (see Prince, 1992; H. H. Clark & Krych, 2004).

In this work, evidence is provided that objects with high visual salience can already be referred to by reduced forms in situated task descriptions, instead of the otherwise typical initial full names even though the reduced form itself is insufficient

to resolve the reference to the referred object (e.g., pronouns without antecedents, or pronouns not matching the gender of the antecedent in German). Although some approaches (e.g., Gundel et al., 2012) adapted their model by including non-verbal cues such as eye gaze and gesture, these models stay very vague on how these non-verbal cues should be integrated.

Also, to be able to resolve content words despite their variation, underspecified noun phrases, disfluency effects, as well as the interpretation of personal pronouns, information transmitted via visual modalities needs to be tightly interlinked with information transmitted via the linguistic modality. In situated communication, visual cues such as poising, exhibiting, deictic gestures, and eye gaze play an important role in referring to objects and actions (H. H. Clark & Krych, 2004). Especially gaze and gestures are often cited as important cues for establishing joint attention (Tomasello & Akhtar, 1995; H. H. Clark & Krych, 2004; Frischen et al., 2007).

2.1.3 Multi-modal cues

Eye gaze. One potential cue for disambiguation is eye gaze (see Prasov & Chai, 2008; Knoeferle & Crocker, 2006; Hanna & Brennan, 2007; H. H. Clark & Krych, 2004). Eye movements are naturally occurring in parallel to speech, they are informative, and they can be used by addressees as visual cues during reference resolution (Hanna & Brennan, 2007). Investigating eye movements during situated utterance production and comprehension has revealed that referential gaze is closely time-locked with the unfolding speech stream (Griffin, 2001; Tanenhaus et al., 1995). Humans' tendency to follow each others' gaze in face-to-face communication is considered to be rather resistant to top-down influences (Böckler et al., 2011).

For resolving ambiguous references, the speaker's gaze in a shared environment also provides listeners with visual cues where the attention of the speaker is focused at (Hanna & Brennan, 2007; H. H. Clark & Krych, 2004). The speakers' eye movements to objects show scanning patterns that reflect the incremental encoding of utterances on conceptual, syntactic, and phonological levels (Griffin, 2001; van der Meulen et al., 2001). Hanna & Brennan (2007) emphasize that gaze can be used communicatively as a form of pointing, to intentionally draw an interlocutor's attention to an object. However, they emphasize that in order to function as a signal, gaze must be integrated with speech or action.

In addition to pronoun resolution, a speaker's eye gaze may function as an indicator for upcoming utterances (Frischen et al., 2007). Research on language production showed that speakers fixate a to-be-named object 800 ms to 1 s prior to the onset of uttering its name (see Meyer et al., 1998; Bock et al., 2003; Rossion &

Pourtois, 2004). Depending on the task, eye gaze behaviour might change. Griffin & Bock (2000) showed that in naming tasks, speech about one object was being produced while the next object was fixated and lexically processed. However, in most of these studies, objects are presented to participants on a screen or a sheet of paper. In situated task descriptions in which participants have to conduct and explain a task, there might be other factors influencing eye gaze of the participants.

Gestures. They are an integral part of language, synchronous and co-expressive with speech, and can be deictic (pointing) gestures, iconic gestures, emblems, and beats (Ekman & Friesen, 1981; Kendon, 2004; McNeill, 2005; Bergmann & Kopp, 2012). In addition, placing things just in the right manner is an indicative act in face-to-face communication similar to pointing (H. H. Clark, 2003). When presenting a task to a learner, deictic gestures as well as the indicative act of placing and manipulating objects are of special interest for resolving referring expressions.

Gestures conducted by speakers can be redundant with the information encoded verbally (e.g., "round cake" + gesture depicting a round shape), supplement (e.g., "cake" + gesture depicting a round shape), or even complement the verbal description (e.g., "looks like this" + gesture depicting a round shape) (Kopp et al., 2013).

Multi-modal communication. In situated task presentations, participants interpret linguistic and visual inferences in parallel. It is thus important, how people divide their efforts and information between vocal and visual actions. H. H. Clark & Krych (2004) conducted a study in which two participants had to collaboratively solve a task: an instructor directed a builder in how to assemble Lego models. In one scenario, the workspace of the builder was visible to the instructor, in another, it was not, and in a third, instructions were given by audiotape. When the workspace was visible, builders communicated with the instructor by exhibiting, poising, pointing at, placing, and orienting blocks, and by eye gaze, head nods, and head shakes, all timed with precision. When the workspace was not visible, the two partners were much slower and in the third scenario, they made more errors. The results provide evidence for the claim that multi-modal information is essential in situated task descriptions. H. H. Clark & Brennan (1991) argue that people are opportunistic in trying to select from the available methods – verbal and non-verbal – the ones they think take the least effort for jointly the speaker and the listener.

Depending on the communication context, speaking might also impede understanding. H. H. Clark & Krych (2004) argue that participants use verbal and non-verbal modalities in parallel and that for certain types of communication the

visual modality is faster and more reliable than the auditory modality. Brennan et al. (2008) conducted a study in which participants had to undertake a search task (i) alone, or in pairs with (ii) shared gaze, (iii) voice, or (iv) shared gaze and voice. Collaborating pairs performed better than solitary participants, but pairs were able so solve the task faster in the shared gaze search than in the shared gaze and voice search. In a study by Lozano & Tversky (2006), communicators explained how to assemble a simple object using either speech with gestures or only gestures. In the "gestures only" - condition, the assembly task was learned better and fewer assembly errors were made than in the "speech with gesture" - condition.

These studies indicate that in some cases verbal descriptions might also impede understanding or negatively impact task-oriented information transmission. Thus, for the design of artificial agents it is crucial to include capabilities for detecting and integrating non-verbal cues in order to supplement verbal descriptions.

2.2 The Interlocutor as the *Immersed Experiencer*

In order to develop mechanisms for robot architectures to deal with situated, task-based interactions, it is not only important to implement mechanisms for identifying and processing multi-modal cues of the interlocutors, but also to account for findings in embodied human language comprehension.

In the last decades, an increasing body of work in psychology and cognitive science has raised evidence for the tight integration of human language processing with sensory and motor-driven experiences (see Barsalou, 2010; A. M. Glenberg & Gallese, 2012; A. M. Glenberg et al., 2013; Pickering & Garrod, 2013; Zwaan, 2014).

A theoretical framework trying to incorporate important studies central to language understanding and developed as a basis for an embodied theory of language comprehension is the Immersed Experiencer Framework (IEF, Zwaan (2004)). It accounts for the following findings:

(i) *The processing of words activates brain regions that are close to or overlap with brain areas that are active during acting or perceiving the words' referents.* Neuroimaging studies have shown that tool words activate motor areas in the brain and words of certain animals activate visual areas (Martin & Chao, 2001; Kiefer & Barsalou, 2011). Similarly, in an experiment by Simmons et al. (2005) pictures of appetizing foods activated gustatory cortices for taste and reward.

(ii) Behavioural experiments revealed *the importance of action in language com-*

prehension. In an experiment by A. Glenberg & Kaschak (2002), participants had to declare whether or not a sentence was meaningful by moving a button. Responses were facilitated, if the movement described in the sentence was in the same direction as the movement of the button: e.g., "He closed the drawer" as a movement away from the subject and "He opened the drawer" as a movement towards the subject. A similar effect was observed for more abstract sentences, e.g., "I told him a story". In a similar experiment by Zwaan & Taylor (2006), sentences with clockwise and counter-clockwise movements were presented to the participants, e.g., "Jane started the car" or "Liza opened the pickle jar". The results show that the responses were faster when the movement to respond was in the same direction than when it was in the opposite direction.

(iii) *Perceptual representations are routinely activated during comprehension.* Stanfield & Zwaan (2001) presented sentences to subjects, e.g., "He pounded the nail into the wall" and "He pounded the nail into the floor", and pictures with a horizontal and a vertical nail. Subjects responded faster when the object orientation of the nail in the picture and in the sentence was the same.

(iv) When humans comprehend language, their *eye and hand movements are consistent with perceiving and acting in the situation described.* Entities, features and objects, ongoing events and current goals that are currently in working memory are more accessible than absent, distant or past ones, see for example (Kaup & Zwaan, 2003; Horton & Rapp, 2003; Rinck & Bower, 2000).

The IEF is a theoretical account to embodied language comprehension, distinguishing three processes: ACTIVATION, CONSTRUAL, and INTEGRATION. ACTIVATION refers to the mental activation of multi-modal representations that are connected to objects and events triggered by the stream of words in an utterance. CONSTRUAL is the (sequential) integration of several functional webs in a mental simulation of an event. Linguistically, the information is encoded at the level of clause and intonation units. INTEGRATION refers to the transition of one construal to the next one. The comprehender proceeds from event representation to event representation, and relevant components of the previous construal influence the current construal, (see Zwaan, 2004; Zwaan & Madden, 2005).

Although the IEF is still being developed, it is to my knowledge currently the only framework considering sensory and motor-driven experiences in the incremental processing of language. Thus, it might be a potential framework for developing language processing mechanisms for robot architectures.

2.3 Summary and Discussion

This chapter reviewed research literature relevant in the scope of this thesis. In Section 2.1, psycholinguistic studies are reviewed whose results might impose challenges for situated multi-modal human-robot interaction. One area concerns **general inter- and intra-speaker variation with regards to language**:

> *Variation of content words*: Humans differ in lexical choices for one and the same object, named "the vocabulary problem". Although in human-human dialogue, lexical entrainment occurs within one interaction, i.e., the interlocutors came to use the same words. However, in situated task descriptions where one person is describing and conducting the task while the other one is mainly listening, no lexical entrainment can occur. The variation in wording raises a problem for human-robot interaction, as the robot still has to resolve, e.g., a verbal referring expression to an entity in the visual scene. The linguistic and visual salience of potential referents need to be accounted for when resolving references.
>
> *Disfluency effects*: Repairs, abandoned utterances, filler words occur frequently and need to be handled.
>
> *Taken perspective*: Personal pronouns can not automatically be interpreted according to their literal meaning.

Computational approaches to the variation of words, disfluency effects, and perspective taking are reviewed in Section 3.1.

Another challenge in human-human interaction is presented by **verbal referring expressions**. Most approaches assume that there is a direct relationship between the form of a referring expression and the accessibility of the referent. Several models for resolving references are reviewed. Most models do not account for the visual salience of objects, only some also include non-verbal cues, but on a very rudimentary level. Details on the interplay of verbal and non-verbal modalities are missing and they are thus also not sufficient as a basis for implementation.

In the third part of this section, literature on **multi-modal cues** was reviewed. Light was shed on the relevance of *eye gaze* and *gesture* in situated communication. Both are closely time-locked with the unfolding speech stream. In order to deal with multi-modal information transmitted by the interlocutor, it is important to investigate how people divide their efforts between verbal and non-verbal cues. There is empirical evidence, that language might also impede understanding in task-based descriptions, when the workspace, eye gaze, or gestures of the interlocutor are visible.

Literature on computational models of reference resolution including multi-modal cues are presented in Section 3.2.

The second section on **the interlocutor as the immersed experiencer** presented a theoretical framework of embodied language comprehension. The Immersed Experiencer Framework tries to account for important findings in psychology and cognitive sciences, showing that perception and action are tightly interlinked with language processing:

(i) Words activate brain regions that are close to or overlap with brain areas that are active when interacting with the words' referents.
(ii) The importance of action: humans respond fast to utterances, if the movement to respond is in the same direction as in the uttered sentence (e.g., clockwise versus counter-clockwise movements).
(iii) Perceptual representations, such as the orientation of objects, are routinely activated during language comprehension.
(iv) Entities, objects, events, etc. that are currently in working memory are more accessible than distant or past ones.

Due to its empirical basis and the aim to consider sensory and motor-driven experiences in the incremental processing of language as comprehensively as possible, this model was selected to map it on the empirical data of situated task descriptions. The mapping presented in Chapter 4 will be used to investigate the applicability of a model of embodied language comprehension on task descriptions within a shared environment.

Chapter 3

Multi-modal Human-Robot Interaction

The previous chapter provided empirical, theoretical, and psycholinguistic perspectives on aspects of multi-modal human-human interaction that impose challenges to human-robot interaction. This chapter will show how computational approaches investigate these challenges.

This chapter will begin with a review of work concerning general challenges for robots in situated interaction with a human interlocutor in Section 3.1. Subsequently, research with regards to computational multi-modal reference resolution will be presented (Section 3.2). In Section 3.3 on data collection for human-robot interaction, the notion and use of corpora in the field of human-robot interaction is discussed, as well as annotation tools an schemes. The chapter will conclude with a summary and discussion in Section 3.4.

3.1 General Challenges for Robots in Situated Human-Robot Task Descriptions

For more natural human-robot interaction, robot architectures must be developed, which enable robots to process linguistic and visual input in parallel in order to extract all information necessary for understanding. In more recent approaches, attempts have been made to also account for multi-modal interaction where the robot for example is able to use gestures to resolve ambiguous references (see McGuire et al., 2002; Mavridis & Roy, 2006; Mavridis, 2007; Dzifcak et al., 2009).

However, due to the complexity of speech recognition in human-robot interaction, the conversational competencies in most current applications are restricted to

a set of motor commands. Mavridis (2015) argues that interaction based on simple motor command requests has the following drawbacks: (i) the dialogue is primarily single-initiative, as the human drives the conversation and the robot produces motor and verbal responses, (ii) the robot does not speak except for some disambiguating questions, (iii) the only occurring speech acts (see Searle, 1969) are requests for motor action, (iv) these systems are usually quite inflexible regarding the surface realisations of the commands, elliptical utterances such as "the red object please" might be misinterpreted, and (v) the mapping of words-to-responses is in most cases arbitrarily chosen by the designer, i.e., motor verbs are not interpreted according to their empirical meaning (based on empirical investigations), but to their normative meaning (what the designer thinks they should mean). Additionally, robots should not require humans to adapt to them in a special way, and be able to fluidly collaborate with humans, interacting with them and being taught by them in a natural manner. Based on this aim and the above listed drawbacks, Mavridis proposes ten desiderata to discuss what capabilities are needed for future human-robot interaction (see Mavridis (2015) for details):

- Breaking the "simple commands only" barrier
- Multiple speech acts
- Mixed initiative dialogue
- Situated language and the establishing representations
- Affective interaction
- Motor correlates and non-verbal communication
- Purposeful speech and planning
- Multi-level learning
- Utilization of online resources and services
- Miscellaneous abilities, such as multiple conversational partners, multilingual capabilities

All of these points are rather large and important areas of research on their own. For the presented work, there is no focus on different kinds of speech acts and mixed initiative dialogue as the task descriptions are primarily monologues with sporadic backchannels from the listeners. For observing and listening in order to extract information for how to conduct a certain task, affective interaction is also not decisive. Still, negative or positive emotions might be reflected in speech, if the task does not work the way intended by the instructor. However, this aspect of the interaction is not necessary for conducting the task. The focus of this thesis is on extracting information transmitted via different channels and the variation in

information. Purposeful speech and planning as well as motor correlates would be relevant for the subsequent step to be able to reproduce a certain task. Likewise, multi-level learning as well as the utilization of online resources, multiple conversational partners and multilingual capabilities are important for subsequent steps, but not for extracting, merging, and interpreting information in situated task descriptions. Of relevance for the current work are situated language, establishing representations and especially non-verbal communication.

For dealing with situated task descriptions, robots have to be able to:

- share representations of concepts (objects, actions, agents, spatio-temporal relations, etc.) underlying the interaction with their human interaction partner, and
- deal with the multi-modal complexity of information while interacting with their human communication partner, i.e., to identify human communicative cues and extract and merge information transmitted via different channels.

In this work, the focus is not on establishing representations, but it is assumed that the representations are already empirically established. Still, the following aspects related to situated language are relevant:

Personal pronouns. These pronouns need to be grounded and interpreted in human-robot interaction. First, robots need to learn the interpretation of "you" and "I". Empirical studies have shown that humans learn "I" and "you" better when they observe others using it (Oshima-Takane et al., 1996). Oshima-Takane et al. (1999) developed a neural network for learning these pronouns. The results with regard to learning speed and analysis of their knowledge representations confirmed the importance of exposure to overheard speech.

Roy et al. (2004) present a set of representations and procedures that enable a robot to maintain a "mental model" of its physical environment by coupling active vision to physical simulation. Within this model, "imagined" views can be generated from arbitrary perspectives (e.g., *my left* versus *your left*).

Gold & Scassellati (2006) developed a system that can learn the correct deictic meaning for "I" and "you" by observing interactions between other agents. It uses contextual information from already understood words and sensory information from its environment. The system also serves as a model for the phenomenon of *pronoun reversal* among congenitally blind children before the age of 5 (Andersen et al., 1984). *Pronoun reversal* is the usage of "you" or another personal pronoun where "I" is meant, or vice versa.

Empirical studies have shown that in many languages impersonal pronouns can be used to transmit structural knowledge and general truths (see Kitagawa & Lehrer, 1990). Thus, there is an additional challenge for interpreting personal pronouns: they might not be used literally, for example if a person is explaining and conducting a task to someone, there are several options which personal pronoun is uttered, e.g., "I/you/we now take ..." although it is always the person talking who is – at the moment – conducting the task.

Meaning negotiation. To resolve what is uttered to what it refers to is also an important task in human-robot interaction. It is often adjusted online during the conversation (see Brennan & Clark, 1996) and the classifier of the human and the interlocutor might not always match (see Mavridis, 2007). For example, if a category is uttered in a situated task description that is not existent in the model of the listener, but there is an object of a similar category, the listener then has to temporarily adjust his/her model online.

The interpretation of personal pronouns and the negotiation of meaning touch the topic of situated language. For a more general overview on situated human-robot interaction see Kruijff et al. (2010); Coradeschi et al. (2013).

Handling ungrammatical or partial utterances. In human natural speech production, abandoned utterances, filler words or repairs frequently occur (see Foxtree & Clark, 1997). Scheutz et al. (2007, 2013) developed the robot architecture DIARC (short for "Distributed Integrated Affect, Reflection, and Cognition" architecture) aiming at more natural human-robot interactions. The architecture includes mechanisms for natural language processing, intentional behaviours, and monitoring mechanisms to detect faults and recover from them. Another approach for handling ungrammatical or partial utterances was proposed by Hüwel et al. (2006). Their model includes speech and gesture input, as well as knowledge about the contextual scene.

The appearance and morphology of the robot. The robot's appearance and the resulting impact on human-robot interaction have been directed in several studies. Parameters influencing and interfering with the interaction are difficult to control and they differ whether the robot has humanoid, child- or adult-like appearance, depend on the degrees of freedom the robot has, and how reactive its behaviour is (see Vollmer et al., 2009; Pitsch et al., 2012). There is also a growing body of literature on impact of robots combining visual and linguistic references in shared scenes on their human interlocutor (see Kranstedt et al., 2006; Van der Sluis & Krahmer,

2007; Staudte & Crocker, 2009; Fang et al., 2015). In this work, the multi-modal variation the listener has to deal with during a situated task explanation is explored. However, the influence of the morphology of a robot needs to be kept in mind whenever investigating human-robot interaction.

3.2 Computational Approaches to Multi-modal Reference Resolution

Over the past two decades, the research area of computational linguistics hast grown and technology developed in this field is increasingly incorporated into consumer products (see Hirschberg & Manning, 2015, for an overview). The authors call the following advances to account: (i) the increase in computing power, (ii) the availability of large amounts of linguistic data, (iii) the development of successful machine learning methods, and (iv) a richer understanding of the structure as well as social contexts of human language.

However, in most approaches for reference resolution, the discourse model is handled separately from the interaction model. In situated task-descriptions, a chief obstacle to developing conversational agents is to develop mechanisms for understanding and producing referring expressions appropriately within the setting. Requirements to computationally interpret and produce referring expressions in shared environments include

(i) a multi-modal knowledge representation containing visual and lexical entries of all entities that are available for reference (e.g., Kruijff et al., 2010, 2006; Coradeschi et al., 2013),
(ii) judgements about the relative salience of entities (e.g., Grosz et al., 1995; Goudbeek & Krahmer, 2012),
(iii) a model of common ground in order to determine how to refer to objects (e.g., Chai et al., 2014),
(iv) recognition of speakers's gaze and gestures and identification of where they are directed at (e.g., Van der Sluis & Krahmer, 2007; Lemaignan et al., 2012).

Only in the last few years have attempts been made to combine experimental research on referring expressions in psycholinguistics and computational work on algorithms that identify and generate referring expressions (Van Deemter et al., 2012; Gatt et al., 2014). A well known computational approach to model anaphoric reference is Centering Theory (Grosz et al., 1995). In Centering Theory, anaphoric references between consecutive utterances have a backward looking centre and a set

of forward looking centres each. The forward looking centres within an utterance are ranked according to their salience. The backward looking centre is the forward looking centre from the previous utterance with the highest rank. In the following example, "John" is the backward looking centre and needs to be pronominalised: "John went to his favourite music store to buy a piano. He had frequented the store for many years." (see Gatt et al., 2014, p.7).

The Incremental Algorithm developed by Dale & Reiter (1995) is especially important for the research area of "Referring Expression Generation" (REG). It tackles the selection of content for a descriptive referential noun phrase and is based on two contradictory developments: (i) according to Grice's Maxim of Quantity, human interlocutors attempt to produce referring expressions that convey no more information than required (H. Grice, 1975); (ii) however, psycholinguistic studies have shown, that humans tend to overspecify referents (see Pechmann, 1989). This tendency includes properties, such as shape, size or colour (e.g., Arts et al., 2011; Goudbeek & Krahmer, 2012).

Krahmer & Theune (2002) have proposed an extension of the Incremental Algorithm, incorporating ideas on how to handle anaphora from Centering Theory. They propose to compute salience of referring expressions based on grammatical role as in Centering Theory. Their extension takes context into account, as pronouns are only generated in case the entity referred to is the most salient entity in the discourse, not only the preceding utterance.

However, these approaches assume that all pronouns can be resolved via antecedents and that referring expressions are not underspecified. In a study by Kowadlo et al. (2010) a spoken language understanding system performed better when no pointing was used by the speaker than when pointing was used, as the speaker uttered more precise referring expressions without gestures. Humans, however, naturally employ these cues and for a robot to be able to resolve these references, a deeper understanding of how they can be identified and interpreted is necessary.

In addition to linguistic referential expressions, some approaches also take into account visual references such as deictic gestures and eye gaze. Ideally, underspecified verbal referring expressions and visual references identify the same object at the same time and thus can still be resolved. Admoni et al. (2014) studied the effects of conflict in human-human and human-robot interaction. Their results show that congruent gaze helps performance in HH and HRI, while incongruent gaze resulted in no longer response times than absent gaze.

Kelleher & Kruijff (2006) developed an extension of the Incremental Algorithm

which adds a notion of visual and discourse salience in addition to contextually defining the set of objects that may function as a landmark.

Gundel et al. (2006) have also proposed a "coding protocol" for the Givenness Hierarchy (see above), assigning different pieces of information to different cognitive statuses, for example targets of gesture or eye gaze are automatically activated, and the syntactic topic of the preceding sentence is assumed to be in focus, thus including information coming from one's dialogue, environmental, and pre-existing knowledge.

In the research area of Human-Robot or Human-Agent interaction several attempts have been made to implement adapted versions of the Givenness Hierarchy. Kehler (2000) proposed an adapted version of the Givenness Hierarchy aiming at resolving multi-modal references in the context of pen-and-tablet interfaces. They applied four simple rules to resolve references: (i) If an object is gestured to, choose that object. (ii) Otherwise, if the currently selected object meets all semantic type constraints imposed by the referring expression, choose that object. (iii) Otherwise, if there is a visible object that is semantically compatible, then choose that object. (iv) Otherwise, a full NP was used that uniquely identified the referent.

Chai et al. (2006) applied a greedy algorithm for combining the Givenness Hierarchy with Conversational Implicature by H. Grice (1975). By combining these two, they derived the following modified hierarchy: *gesture* > *focus* (subsuming Gundel's *in focus* and *activated* tiers) > *visible* (subsuming Gundel's *activated* and *uniquely identifiable* tiers) > *others* (subsuming Gundel's *referential* and *type identifiable* tiers). Their greedy algorithm is able to handle ambiguities and multiple references in one utterance.

Williams et al. (2015) propose an implementation for the Givenness Hierarchy handling definite and indefinite noun phrases, and pronominal expressions, thus allowing the algorithm to deal with a wider range of linguistic expressions than previous approaches, see also Chapter 7. The presented algorithm is also able to handle open world and uncertain contexts, though it has not yet been evaluated on a robot.

Besides the Incremental Algorithm and the Givenness Hierarchy, there are also other approaches dealing with multi-modal reference resolution. Prasov & Chai (2008) developed a probabilistic framework to combine linguistic referential expressions and eye gaze to decrease the need for a complex pre-defined domain model to resolve referring expressions.

Implications for the inherent combination of visual and linguistic references in shared scenes were investigated by Staudte & Crocker (2009). In their study, robot gaze had an even stronger influence on people's visual attention than other linguistic

cues.

Van der Sluis & Krahmer (2007) developed a graph-based algorithm for generating deictic pointing in combination with linguistic referring expressions. In their approach, costs are assigned to linguistic properties and pointing gestures in the generation of multi-modal referring expressions. In another study investigating the usage of pointing gestures and linguistic referring expressions, Kranstedt et al. (2006) aim to model the focussed area of pointing gestures (the "pointing cone") in combination with verbal references.

Lemaignan et al. (2012) propose an approach to extract, represent, and use knowledge from real-world perception as well as from human-robot verbal and non-verbal interaction. Strategies for disambiguating concepts include whether the previous interaction involved a specific action and whether the user is looking or pointing at a specific object. Their current implementation relies on a small, predefined set of action verbs that can be recognized from natural language.

Chai et al. (2014) emphasize that although humans and robots are situated in a shared environment, internal representations of objects are misaligned. They developed a model for common ground in situated human-robot dialogue in order to still be able to resolve references to objects. A reference resolver matches a dialogue graph to a vision graph (capturing the robot's internal representation of the shared environment), applying an inexact graph-matching algorithm. Fang et al. (2015) employ the same algorithm to generate referring expressions. In addition to verbal references, they include deictic gestures by the robot and human's gaze feedback. However, they focus on whether the human is able to resolve the referring expressions generated by the robot and not the other way around.

Huang & Mutlu (2014) develop a dynamic Bayesian network (DBN) for modelling how humans coordinate speech, gaze, and gesture behaviour in narration, learn model parameters from annotated data, and draw on the learned model to coordinate these modalities on a robot.

There is also a growing body of literature on how humans differently react to robots that combine visual and linguistic references in shared scenes (see Kranstedt et al., 2006; Van der Sluis & Krahmer, 2007; Staudte & Crocker, 2009; Fang et al., 2015). However, this is beyond the scope of this thesis.

3.3 Collecting Data for Human-Robot Interaction

Situated interaction corpora are invaluable resources for investigating the complex relationships among language, perception, and action. In order to make use of

empirical, psycholinguistic data to develop computational models, data collections or corpora are needed.

In this section, first, the notion of "corpus" is introduced. Subsequently, the use of data collections in the field of human-robot interaction is discussed. The section is concluded with an overview on different annotation tools.

3.3.1 The notion of "corpus"

Depending on the research area, the usage of the term "corpus" differs. In corpus linguistics, corpora can be generally defined as "a body of naturally occurring language" (McEnery et al., 2006, p.4), whereby the interpretation of *naturally* is crucial. Gries (2009) (p.8) argues that "the texts that make up the corpus must have been produced in a natural communicative setting. That means that the texts were spoken or written for some authentic communicative purpose, but not for the purpose of putting them into a corpus". Two other criteria are *representativeness* of the corpus for a certain language, variety, or register and that the corpus is *balanced* (the size of the parts of the corpus corresponds to the proportion these parts make up in the language/variety/register) (Gries & Newman, 2013). Gries & Newman (2013) state that large corpora contain about 100 million words or more and they give the example of a relatively small corpus by Berkenfield (2001) containing 10,640 words.

When working with empirical data, two traditions can be distinguished: "corpus-based" and "corpus-driven" approaches (see Hardie & McEnery, 2010). In corpus-based approaches, corpus techniques can be applied in different fields of language study, while representatives of the corpus-driven camp argue that there is no role in corpus linguistics for theories of language that do not emerge from the study of corpus data (see Teubert, 2005). Teubert (2005) (p.2-3) also argues that corpus linguistics "is not concerned with the psychological aspects of language". The research conducted within this work is a data-driven, qualitative approach as relevant aspects are inductively extracted and later combined with a quantitative analysis. Nevertheless, it does not fit within the corpus-driven camp, as psychological aspects and theories of human-human communication play a crucial role for the analysis. Lindquist (2009) denotes an approach as "corpus-driven" if as few preconceived theoretical concepts as possible are used at the beginning, "corpus-based" if quantitative methods are used to investigate a problem which is already formulated within a linguistic theory, and "corpus-aided" if corpora are used to find illustrative examples. Lindquist's definition of "corpus-driven" fits better to the approach used in this dissertation.

In the area of human-robot interaction and computational modelling, data collections are usually smaller and are often still referred to as "corpora". For example, the corpus built by Gaspers et al. (2014) contains task descriptions by 27 participants, about 20-30 minutes per participant, in the corpus by Green et al. (2006) 22 users interact for about 15 minutes each, or the Vernissage corpus by Jayagopi et al. (2013) comprises interactions from 26 participants with an average of 60 utterances per participant. A reason for keeping multi-modal corpora rather small is the effort needed for preparing the data. Aim of these corpora is to develop computational models for natural human-robot interaction. The results of these studies, however, can also be used to inform research on human-human interaction. Research in this area focuses more and more on multi-modal, spontaneous spoken interaction in a shared environment (see Abuczki & Ghazaleh (2013) or Tenbrink et al. (2013) for an overview on multi-modal corpora, annotation tools and schemes). *Natural* data in this context means that the human does not have to adapt to the robot by using a command-like language but by transmitting the relevant information in a way natural to humans. Newspaper articles written for some authentic communication purpose are less natural for human-robot interaction than utterances spontaneously produced in a shared environment collected for the purpose of putting them in a corpus. However, in order to avoid conflicts between research areas due to different foci of the disciplines, the data collected for this thesis are referred to as data collection not as corpus.

3.3.2 Data collections in the field of human-robot interaction

Corpus data can be used to better understand the intricate processes involved in human-human situated interaction. This is not only central to a better understanding of human natural language interactions, but also critical for research in human-robot interaction (e.g., Green et al., 2006; Rehm & André, 2008; Tenbrink et al., 2013). Data on human-human instructor-learner dyads are a valuable resource to develop computational models for robots.

There are several multi-modal annotation schemes available and they need to be chosen according to the level of granularity which is relevant for the research question. These decisions need to be made for each annotation tier. The tier also needs to be chosen according to the research questions at hand. Potential layers are:

Dialogue acts: Annotation schemes focusing on dialogue acts include e.g., the DIT++ (Bunt, 2009), DAMSL (Allen & Core, 1997), or the HCRC coding

scheme (Carletta et al., 1997).

Intonation: This aspect of situated interaction is covered e.g., by ToBI (Silverman et al., 1992), the German version GTobi (M. Grice & Baumann, 2002), or DIMA (Kügler et al., 2015).

Gestures: Schemes for gesture coding include e.g., FORM (Martell et al., 2002) or the very detailed MUMIN coding scheme (Allwood et al., 2007).

In most cases it is unfavourable to adopt a coding scheme as is, but to choose aspects of the schemes relevant for the research questions.

The majority of instructor-learner corpora are caregiver-child interactions. A large resource is the CHILDES data base which serves as a central repository for first language acquisition data. Moreover, Björkenstam and Wirén (2013), as well as Yu et al. (2008) collected and annotated multi-modal caregiver-child interactions. In this data, the interactions are spontaneous and the tasks not pre-defined. In this thesis, the focus is on the variation of the communication signals between, but also within task descriptions. Thus, different people need to explain the same task, in order to shed light on the variation how humans naturally structure and present information.

Another valuable source are thus task-oriented interactions. The MAP Task (Anderson et al., 1991) is a widely used paradigm that has been employed in different versions. It involves two participants who both have not-identical maps (and are informed of this), which the other person cannot see. The maps contain drawings of labelled landmarks (e.g., rocks, bridge, mountain, etc.). The HCRC (Human Communication Research Centre) coding scheme was originally developed for the MAP Task and the TRAINS corpus. According to HCRC, the three levels of dialogue acts "transaction", "game", and "move levels" were annotated.

Another well known corpus is the TRAINS corpus (Heeman & Allen, 1994, 1995) consisting of dialogues between two participant in a problem-solving task. Goods need to be shipped in various cities by trains. DAMSL was used as an annotation scheme, using communication management, task management, and task levels for the annotation of dialogue acts.

In other data collections, participants had to build toy airplanes from parts (Rickheit & Wachsmuth, 2006). The setting involved sufficient complexity of actions to involve a high amount of negotiation between the instructor-constructor dyads. Participants were separated by a screen and could thus not see each other. The authors are interested in the aspects of the interaction that will lead to shared

common ground and emphasize the importance of prosodic features for the update of common ground.

In the Dollhouse scenario (Tenbrink et al., 2008), pairs of participants were asked to furnish a dollhouse. The setup allows for the less informed person ("the matcher") to contribute to the continuous update of common ground by making relevant suggestions to the director. The directors had full visual information about positions of objects but not of the workplace of the matchers. The matchers on the other hand had to place objects into an empty dollhouse based on the directors' instructions.

The CReST corpus (Indiana Cooperative Remote Search Task) (Eberhard et al., 2010) comprises 16 dyads performing a cooperative search task. The director was located in a room distant from the search environment, directing the searcher through the environment by a telephone. The results of the analysis show the importance of dialogue for updating common ground and coordinating joint actions in a remote scenario.

In a multi-modal interaction study by Anastasiou (2012), the interaction between a powered wheelchair called Rolland and a user who was asked to carry out a set of tasks with the wheelchair was investigated. A Wizard-of-Oz setting was employed and the focus of the annotation lay on the utterances, actions, gestures, and dialogue acts. Although the interaction capabilities of the wheelchair are clearly limited, the analysis of the corpus illustrates the important role of gesture in human-robot interaction.

In the German corpus by Gaspers et al. (2014) 27 participants prepared dishes in front of an iCub, using toy objects. A small set of actions and objects was selected for the tasks and audio, video, as well as motion was recorded using the Kinect. The corpus comprises multi-modal data which support the evaluation of computational models of multi-modal language processing, with a focus on learning tasks for language acquisition in robots.

3.3.3 Annotation tools

There exist a variety of different annotation tools with different advantages and disadvantages. Depending on the research question and needs of the data to be analysed, an annotation tool has to be chosen, e.g., not all tools allow for all audio and video formats.

The following tools are among the most commonly used annotation tools for spoken language transcription:

- ANVIL is a multi-layered video annotation tool originally developed for studies of multi-modal behaviour (Kipp, 2001) (`http://www.anvil-software.org/`).
- ELAN is an audio and video annotation tool which can be used for example for sign-language transcriptions (Kipp, 2001) (`https://tla.mpi.nl/tools/tla-tools/elan/`).
- EXMARaLDA ("Extensible Markup Language for Discourse Annotation") consists of a transcription and annotation tool (Partitur-Editor), a tool for managing corpora (Corpus-Manager) and a query and analysis tool. It was originally developed for the analysis of multilingual data (Schmidt & Wörner, 2009) (`http://www.exmaralda.org/`).
- CLAN has an editor which can be used to edit files in either CHAT or CA (Conversation Analysis) format and it's second part is a set of data analysis programs. It is part of the CHILDES database and developed for analysing conversational interaction, language learning, or language disorder (MacWhinney, 2000) (`http://childes.psy.cmu.edu/`).
- FOLKER is a transcription editor originally developed for analysing and searching the FOLK corpus (a research and teaching corpus) for conversation analysis (Schmidt & Schütte, 2010) (`http://agd.ids-mannheim.de/folker.shtml`).
- Praat is a software for phonetic analysis (Boersma, 2002) (`http://www.fon.hum.uva.nl/praat/`).
- Transcriber is an editor built on the EXMARaLDA technology and focuses on the manual annotation of speech signals (e.g., segmenting long duration speech recordings, transcribing them, and labelling speech turns, topic changes and acoustic conditions). The editor was originally developed for transcription of broadcast news (Barras et al., 2001) (`http://trans.sourceforge.net/en/presentation.php`.

3.4 Summary and Discussion

This chapter aimed to cover aspects and challenges of situated, multi-modal human-robot interaction.

In Section 3.1, **general challenges for situated human-robot interaction** were presented. In line with the psycholinguistic studies reviewed in Section 2.1, computational approaches to the variation of words, disfluency effects, and perspective taking were discussed:

Personal pronouns: Both learning and resolution of personal pronouns pose a

challenge for robot architectures. A study on language acquisition in children as well as the evaluation of a computational model on the learning of personal pronouns suggest that "you" and "I" are learned faster when observing others using them. In addition, it is not always adequate to interpret these pronouns literally in situated interactions.

Meaning negotiation: Listeners have to continuously and temporarily update their internal model, if their interlocutors utter categories, that are not in the model of the listener, but only a similar category (e.g., "tube" versus "pipe").

Handling ungrammatical utterances: In human situated communication, utterances contain filler words, corrections, are ungrammatical or fragmentary. Some attempts have been made to deal with these utterances from a computational perspective.

Appearance of the robot: In addition, several studies have shown that the appearance and morphology of the robot (e.g., the degrees-of-freedom of its limbs) influence how the human perceives and interacts with the robot.

With regards to situated language and shared representations, this thesis argues for the importance, that the meaning of utterances are based on empirical observations and can not be hand-crafted by the designer. In situated human-robot task descriptions on the one hand (i) the utterances are "situated" such that they refer to the physical here-and-now, and on the other hand (ii) both verbal and non-verbal information need to be integrated for fully grasping the information necessary to be able to conduct the task.

In Section 3.2, literature on **computational models of reference resolution including multi-modal cues** were reviewed. Although previous work has presented bits and pieces of people's verbal and non-verbal referring behaviour in inherently multi-modal situated communication, I am not aware of a study as comprehensive as the one presented in this thesis. Some computational models of reference resolution include gestures and / or eye gaze of the interlocutor. However, information transmitted via utterances, eye gaze, and gestures might not be sufficient to resolve all referring expressions in situated task descriptions. Hence, it is critical that we develop more comprehensive computational models of human reference resolution in task-based contexts where instructor and instructee are co-located. Accounting for non-verbal communicative cues is not simply an "add-on" to language processing, but rather an integrative part. Hence, both verbal and non-verbal processing need to be handled flexibly and might contribute essential

information for reference resolution. This will not only inform the theory of situated natural language interactions, but also provide important design principles and constraints for the development of artificial agents that interact with humans in such contexts.

Interaction patterns in situated task scenarios differ substantially from those identified in purely language-based interaction setting. Corpus data is a valuable resource to better understand the relevant processes involved in human-human situated task descriptions and thus, to develop computational models for robots.

In Section 3.3, first, a notion of "corpus" was introduced. Subsequently, a selection of corpora in the field of human-robot interaction was introduced, as well as commonly used annotation schemes. In most cases, coding schemes need to be adapted to the accordant research questions. The drawback of teacher-learner scenarios is that they are not comparable between tasks and most task-based corpora focusing on joint actions are remote and not within a shared environment. In order to develop computational models for robots to deal with human multi-modal complexity, corpora with richer annotations are required, more non-verbal cues than eye gaze and gesture.

Multi-modal layers of interaction were partly taken into account, however, no annotation scheme has been proposed that systematically captures the structural integration of verbal and non-verbal dialogue contribution, even though this is a necessary prerequisite to investigate reference resolution.

The chapter concluded with an overview on annotation tools.

Chapter 4

Pilot Study

Before developing the setup for the data collection, audio and video data already collected by researchers at ITR[1] at the Technical University of Munich were analysed. The corpus comprises 19 German recordings (video plus audio) where one person shows to a learner how to mount a tube in a box with holdings. The instructor performs the task and verbally explains what has to be done. The learner is told to carefully watch and listen to be able to pass the information on to a new learner. The utterances of the instructors and a frontal video of the setting including arms, hands, and torso of the teacher and the learner are recorded. Unfortunately, eye gaze of the teacher is not visible on the videos. The pilot study served as groundwork for developing the more extensive data collection capturing further cues necessary for understanding, including 3 videos (a close-up of the setting, of the teacher, and the learner), motion data of the teacher, as well as force data during collaborative object manipulation (see Section 5).

In Section 4.1, the data collection is introduced. The task itself is very similar to Task 3 of the data collected for this thesis. As a first step, a theoretical model of embodied language understanding (the Immersed Experiencer Framework by Zwaan (2004)) is employed on the data in order to investigate its applicability on situated task description (Section 4.2). An annotation scheme will be developed, accounting both for the Immersed Experiencer Framework as well as for characteristics of situated task descriptions, extracted from the empirical data. The chapter will conclude with a summary and discussion in Section 4.3.

[1] https://www.itr.ei.tum.de/

4.1 Data Collection

The first data collection was used to exploratively investigate multi-modal human-human task descriptions in a shared environment and to draw first conclusions which phenomena a robot would have to deal with, if it were in the learner's position. The participants were students from the Technical University Munich (16 male, 3 female) with German as their mother tongue. The data comprises 19 German recordings (video plus audio) where an instructor shows to a naïve learner how to mount a tube in a box with holdings, see Figure 4.1. Two markers differing in colour have to be put in two different pair of green holdings. The instructor performs the task and verbally explains what has to be done. The learner is told to carefully watch and listen to be able to become the new instructor and pass the information on to a new learner. Thus, the data contain language mirroring the human perception and structuring of the task and its setting. On average, the task duration was 21 seconds (12-34 sec). A frontal video of the setting was recorded including arms, hands, and torso of the instructor and the learner as well as the utterances of the instructor via a wireless microphone.

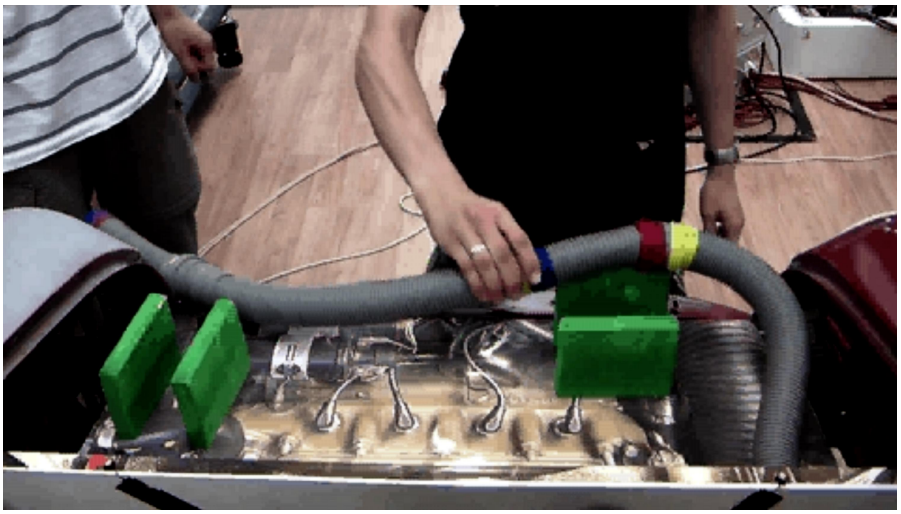

Figure 4.1: A picture of the setting. An instructor is mounting a tube in a box with holdings.

The data were recorded and analysed in order to derive first insight in simple, situated human-human task description. The analyses of the data are the initial step for developing a more extensive setup for the second data collection. Demonstrating and explaining a task in parallel is a valuable source from which insights can be gained on how attention is guided, information is structured, and how the

transmission of information can be achieved through the interplay of different communication channels.

ANVIL[2] was employed for synchronisation and annotation purposes.

4.2 Employing a Cognitive Framework of Incremental Language Understanding for Data Analysis

Although the IEF by Zwaan (2004) is still being developed, it is to my knowledge currently the only framework considering sensory and motor-driven experiences in the incremental processing of language. Thus, it might be a potential framework for developing language processing mechanisms for robot architectures.

As a first step, annotation guidelines were developed which account for the theoretical insights from the IEF and combined with representations from linguistic analysis, information structure, non-verbal communicative behaviour, and low-level signals from the robot's perception and motor systems, based on the collected data.

The aim of this first analysis was to investigate the pros and cons of employing the IEF (developed for non-situated story-telling) for situated task descriptions, and how the IEF can be employed for annotation.

4.2.1 Developing an annotation framework based on the empirical data and an adaptation of the IEF

When applying the IEF on the empirical data, several limitations occurred which need to be dealt with, mainly stemming from the fact that IEF was not developed for situated task descriptions and because it is still at a cursory level. Accordingly, the presented work is the first approach to systematically employ an IEF-inspired view on multi-modal data where a teacher instructs a learner (taking the viewpoint of an immersed experiencer) how to conduct a certain task. Task descriptions were selected because they are an important communication context for human-robot interaction. Demonstrating and explaining actions in parallel is a valuable source from which insights can be derived on how information is structured, attention is guided, and how this is achieved through the interplay of different communication channels.

[2] http://www.anvil-software.de/, (Kipp, 2010)

Based on the analysis of the data, annotation tiers have been defined along the IEF processes ACTIVATION and CONSTRUAL, which are responsible for neural activation triggered by words and activation changes during language understanding. The process of INTEGRATION as the transition from one construal to the next one is reflected in the construals tiers. In the following, the annotation tiers are presented including (i) aspects covered by the IEF and (ii) additional aspects relevant for task manipulation which the IEF does not include or is too unspecific, see Table 4.1.

ACTIVATION

ACTIVATION in the IEF operates on words and is represented by tiers containing (i) the transliteration of the utterances, (ii) an object tier, and (iii) one for actions.

In addition, tiers were added to account for the lack of linguistic detail in the IEF: (iv) a transcript preserving properties of the spoken utterance such as hesitations (e.g., *ahm*, *ah*), contractions (e.g., *dus* (you+it), *gemma* (go+we)) etc., (v) a part-of-speech sequence, and (vi) a representation of syntactic structure, see Table 4.1.

CONSTRUAL

In the IEF, the referential unit of a CONSTRUAL is an event operating on intonation units. Events take place at a certain *time* and in a certain *spatial region*. Within the spatio-temporal framework, there is a *perspective* and within the perspective, there is a *focal entity*, a *relation*, and a *background entity*, each of which may be equipped with specific features. In linguistic research, the correspondence between intonation and linguistic structure is still under discussion (see Büring, 2012). Intonation contour and pauses are often used as indicators for intonation phrases. In the collected data, pauses either break up information for the interlocutor or are indicators for increased cognitive load of the speaker. Thus, CONSTRUAL is structured on the basis of focal entity, relation (verb or preposition), and background entity if one exists, in the developed annotation scheme. Praat is used for analysing pitch contour and intonational phrasing.

In CONSTRUALS, events take place during certain **time intervals**. Zwaan (2004) argues that humans keep events active in working memory trough extended time intervals, as long as the event is ongoing.

In the empirical data, three teachers verbally signalled their respective learner that the task will now start, e.g., "it is about" [...] (*es geht darum* [...])[3], "the goal is" [...] (*Ziel ist* [...]), 10 told their learners when the task was done, e.g., [...] "that

[3]For better readability, the English translation is in the main text and the actual German word choice is in brackets.

was it" ([...] *das wars*), [...] "that's all" ([...] *das ist alles*). All teachers used lexical time markers, such as "first" (*zuerst*), "then" (*dann*), "subsequently" (*anschließend*) to signal the sequencing of the sub-tasks.

The IEF distinguishes between personal space (1.5m around the observer), action space (30m radius), and vista space (beyond 30m), which is too coarse-grained for situated task descriptions. In the annotation scheme, **space** will be encoded by means of the trajectories of the body or body parts of the experiencer during task performance and explanation.

Perspective in the IEF has four aspects: location, distance, orientation, and psychological perspective. *Location* (e.g., verb-induced perspectival changes, such as "X comes into the room") and *distance* (e.g., "molehill" implies a different distance between experiencer and the visual object than "mountain" does) as conceived in the IEF are too high-level for manipulation tasks. Alternatively, a form of location and distance information is encoded via body trajectories of the person explaining and showing the task and by the coordinates or coordinate changes of the objects and agents involved in the task. *Orientation* in the IEF is the physical orientation of the experiencer along the coronal, transverse, and sagittal dimensions.

In manipulation tasks, the orientation or placement (see H. H. Clark, 2003) of objects and agents within the workspace convey more information in order to be able to conduct a task than solely the physical orientation of the experiencer. *Psychological perspective* in the IEF refers to emotions, goals, and knowledge. The context, in which the manipulation tasks take place are relatively neutral, still the emotional perspective could be interesting to investigate. However, this aspect of communication will not be part of this work. And even though the instructor transfers knowledge and the goal of the task, perspective taking is of particular interest in task descriptions.

In the data, 13 teachers used 2^{nd} person singular when explaining while carrying out the task by themselves, e.g., "you grasp the tube with the right hand" (*du greifst den Schlauch mit der rechten Hand*). One participant interpreted an uttered "you" (*du*) as referential "you", and made a step forward to conduct the task himself. When the teacher continued explaining, he stepped back again to watch and listen. Another three teachers used imperative "you have to [...]" (*du musst* [...]). Elliptic form – "first to grasp here" (*zuerst hier greifen*), 1^{st} person plural – "we have to insert the tube here" (*wir müssen den Schlauch hier einfädeln*), and 3^{rd} person singular – "Muriel has to..." (*Muriel muss...*) were used by one person each. One teacher who started with 2^{nd} person singular and the teacher who used 1^{st} person plural switched to the elliptical form during explanation.

In addition to the dimensions considered in the IEF, there are several other aspects necessary for understanding utterances from an instructor in situated task descriptions.

Characteristics of spoken language

Several properties typical for spoken language are present in the data: wrong word substitutions – "holdings" (*Hindernis*) instead of "marker" (*Markierung*); repairs – "red aehm blue and yellow marker" (*rot äh blau-gelben Marker*); insertions – "äh"; contractions – "through the" (*durchs*, "durch das"); errors – *habst* for "have" ('hast').

Variations in wording

Instructors varied in how they verbally referred to objects and actions between, but also within task descriptions. Objects relevant for the task are the tube, two pair of holdings, and three markers. For the tube, all teachers used the same German word *Schlauch* (tube), except for three who did not verbally refer to the object at all. For "marker", two teachers used the anglicism *Marker*, and two used either "point" (*Punkt*) and "gripping point" (*Greifpunkt*) or "endpoint" (*Punkt / Endpunkt*). The other 15 teachers used "marker" (*Markierung*). For the holdings, there was a wide variation in naming: "obstacle" (*Hindernis*), "thing" (*Ding*), "block" (*Block*), "beam" (*Balken*), "rail" (*Schiene*), "marker" (*Markierung*), "log" (*Klotz*), "opening" (*Öffnung*). Again, there was one teacher who did not verbally refer to the holdings. The actions "grasping the tube" and "mounting the tube in the box with the holdings" also showed some variance. For grasping, "grasp" (*greifen*), "have" (*haben*), "take" (*nehmen*), "span" (*umfassen*), "change grip" (*umgreifen*) were used, and for "putting the tube between the holdings": "put" (*legen*), "insert" (*führen / einführen / einspannen / einlegen / einsetzen / einfügen*), "put inside" (*reinstellen / reinlegen*), "clamp" (*klemmen*), and "thread" (*einfädeln*).

Taking the above into account, the learner – may it be a human or a robot – has to infer objects and actions by listening in combination with visual cues in order to be able to resolve references. The action is still the same, although 11 different verbs were uttered (up to two per teacher for the same action).

Verbal and gestural references

During face-to-face communication, a multitude of non-verbal behaviours (e.g., head nods, facial expressions, gestures etc.) accompany speech. Bergmann & Kopp (2012)

Table 4.1: Summary of the annotation scheme. Annotation tiers marked with an Asterisk are adopted from the IEF, tiers marked with a triangle are inspired by the IEF. All other tiers are added to account for the multi-modality of situated interaction.

Process/Modality			**Explanation/Tag**
Transcription			Spoken words (incl. contractions, interjections...)
* Transliteration			Orthographic transcription of the utterance
Activation	Grammar	POS	Automatic annotation tool
		Syntactic structure	Automatic annotation tool
	* Object		Name of the object
	* Action		Name of the action
* Pitch			Praat pitch contour
Construal	Time	▷ Time interval	begin, middle, end
		▷ Time marker	words (e.g., *first*)
	* Entity		Background entity, focal entity, relation
▷ Placement			E.g., right hand on blue and red marker
▷ Perspective of the instructor			E.g. 2nd person singular = IE
Eye gaze of the instructor			Where the eye gaze is directed at
Posture of the instructor			Towards scene, listener, scene and listener
Gesture	Communicative gesture		E.g. deictic, iconic + where the gesture is directed at / what it depicts
	Object manipulation		Object manipulation
	Adaptor gesture		E.g. scratching

emphasize that gestures are in form and timing very closely linked to the semantic content of the speech they accompany (see also McNeill, 2005). Gesture and speech are believed to emerge from the same underlying cognitive representations and are (at least partly) governed by the same cognitive processes (Kendon, 2004; McNeill, 2005).

13 teachers verbally referred to objects, actions or locations, e.g., "here" (*hier*), "like this" (*so*), "this obstacle" (*dieses Hindernis*). The most frequent kind of gestures during task explanations were pointing gestures and holds during object ma-

nipulation to refer to objects or actions in the visual scene. Both gestures serve as indicators for directing the attention of the listener to certain objects or actions. Three teachers used verbal references and communicative gestures simultaneously (e.g., "here" (*hier*) + pointing gesture). One teacher neither used communicative gestures nor verbal references to the visual scene. He only mentioned the grasping of the marker and did not mention that the tube has to be mounted in the box with the holdings. This could only be inferred by the learner from the visual scene. In this respect, Herbert Clark argues that "placing things just in the right manner" (H. H. Clark, 2003, p.243) is an indicative act in which an object is moved into the addressee's attention.

A number of coding schemes for non-verbal behaviour exist, some of which are rather extensive e.g., the MUMIN multi-modal scheme for the annotation of multi-modal communicative behaviours by Allwood et al. (2007) and the "body action and posture coding scheme" (BAP) by Dael et al. (2012). The chosen multi-modal coding scheme has to be adapted to the requirements of the data which comprises mainly object manipulation and deictic gestures. Thus, in the presented coding scheme representations for object manipulation, communicative gestures (e.g., pointing gestures, communicative holds during object manipulation) and posture of the instructor (towards scene, listener, scene and listener) are relevant.

The developed annotation scheme incorporates both the IEF as well as multi-modal aspects of situated task descriptions. Based on the research questions in the upcoming analysis of the newly collected data, parts of the annotation scheme will be used for analysis.

4.3 Summary and Discussion

In this chapter, phenomena were discussed occurring in a corpus of 19 simple task descriptions (action plus speech) of how to mount a tube in a box with holdings. They include characteristics of spoken language, variations in wording, verbal time marking, variation of teacher's perspective, and verbal and gestural references to the scene. These results highlight the importance of multi-modal signal processing in human-robot interaction. In the IEF, objects and actions trigger neural activation. With regards to architecture design, high-level representations need to be time-aligned with low-level sensory data from the robot. However, the manner of internal representation within robot architectures is beyond the scope of this work.

First, the **data collected** at the Technical University Munich were described (Section 4.1). Aim of the recordings was to motivate a more extensive data collec-

tion including a larger amount of recordings via a larger amount and diversity of recording devices, and to give first insights in the variety of multi-modal task descriptions. The analysis of the data also showed a variation in wording much larger than expected, putting an emphasis on the integration of non-verbal cues.

The data analysed for the pilot study included one video of the arms, hands, and torso of the teacher and the learner, hence, eye gaze was not visible. Additionally, two cameras were added in order to allow a focus on the instructor, the learner, and the setting, even if they are not that close to each other (depending on the task). Also, motion tracking devices were added, such as the Kinect sensor or the Qualisys Motion Sensor[4] as well as a force sensor for a collaborate task.

The large variation in how participants structured and transmitted information also inspired the setup of the four different tasks. As there was little information transmitted via language, voiced objects were included in Task 1 to allow for a better analysis of information structure. Task 2 included one very salient object, versus Task 3 with many different although similar objects. There was also a collaborative Task 2 versus Task 3 where the instructor was conducting and explaining the task, while the learner was mainly listening and observing. Task 4 included a navigation scenario where the learner was instructed how to move to a certain location. The reason for including these various foci in the tasks was to investigate how these different setups influence the structure of information in order to shed light on general principles underlying the variation in situated task descriptions.

In Section 4.2, the **Immersed Experiencer Framework** was employed to investigate the applicability of general models of embodied language comprehension in humans. This qualitative analysis of the pilot data collection showed that being an *immersed experiencer* lacks aspects of non-verbal situated communication. Information transmitted via non-verbal channels is crucial for resolving references during situated task descriptions. If only the utterances during task descriptions are interpreted, important information for successfully conducting the task is missing. Although the task was quite simple and the learners had the assignment to listen carefully and forward the information to a new learner, there was quite some variation in how teachers presented the task.

On the one hand, this is due to the variation in naming objects and actions within and between tasks. On the other hand, in situated communication, a multitude of non-verbal behaviours (e.g., head nods, facial expressions, gestures etc.) accompany speech. In addition, characteristics of spoken language such as interruptions, corrections, and so on might impede language understanding.

[4] `http://www.qualisys.com/`

These aspects of situated task descriptions are very crucial, in particular with respect to human-robot interaction and highlight several **challenges for robot architectures**. The occurrence of interruptions, corrections etc. call for robust incremental language processing, in addition to standard language technology tools such as automatic speech recognition, tokenization, part-of-speech, morphological analysis, phrase chunking, dependency parsing, and the such. And although instructors varied vastly when referring to one object or action, the robot still has to be able to resolve the connection of an abstract entity to an entity in the world, e.g., the words *Block*, *Klotz*, and *Hindernis* are all three referring to the green holdings. A comparison of what is visually perceived and what is uttered reveals how differently the same actions and objects are verbally expressed. In addition, the omitted and unspoken entities need to be grounded in the visual scene. As Clark and Krych put it: "when the workspace is visible, the partners ground what they say not only with speech, but with gestures and other actions" (H. H. Clark & Krych, 2004, p.69). Thus, even though some teachers did not mention important objects or actions of the task, the interlocutors were able to resolve these references due to their visual salience.

In order to deal with temporal aspects, a (simple) model of before, after, and concurrency along a common timeline is required together with mechanisms to identify and interpret cues for temporal structuring. These may be lexical (as above), grammatical (tense) or determined by the course of multi-modal action. The change in perspective taken by the instructor is also difficult to interpret by the robot. Thus, the following capabilities are required: (i) the ability to distinguish between the perception of self and other, (ii) a robust interpretation of the perspective from which the action accompanying utterance is issued, and (iii) a model for taking initiative, i.e., for the observer to understand when to just go on observing and when to step in the actor's position.

For gestural and verbal references to visual perception, the robot has to be able to deal with (i) object recognition, (ii) feature recognition, and (iii) gesture recognition. In addition to visual gesture recognition and the recognition of verbal reference to visual perception such as "here" (*hier*), "like this" (*so*), (iv) an attention model is required to enable the robot to detect and interpret the attention directing signals issued by the teacher.

Depending on the phenomena and their functional challenge, there exist none up to a variety of proposed technical solutions for the design of robot architectures. The interplay of the components and the requirement for real-time processing are still far from being reached. More research and integration work is needed on the

way toward human-like task-based natural language processing for robots.

For natural human-robot interaction, on the one hand robots have to be able to share representations of action, objects, agents, etc. with their interlocutor, and on the other hand have to identify human communicative cues and extract and merge information transmitted via multi-modal channels.

The IEF framework is a very promising basis for dealing with the first aspect: modelling representation mechanisms, so that an artificial agent can connect natural language signals with its current action and perception space. However, it was not developed to deal with the second aspect, i.e., situated language, and therefore does not cover very important aspects for situated task descriptions. It is thus not sufficient to deal with this interaction-context, which this thesis is focused on.

The developed annotation scheme tries to be as extensive as possible and reflects potential aspects of situated task descriptions. In the subsequent chapter, a subset of the above presented annotation tiers will be employed, depending on the research questions. However, in future or related work, another subset of the above presented tiers might be useful for annotation.

Chapter 5

Empirical Study on Human Multi-modal Task Descriptions

In this chapter, first, the data collection is introduced (Section 5.1). Subsequently, three different analyses of the data are presented. The first one is an explorative approach, based on the results of the pilot study (Section 5.2). A major challenge in the data is the variation of wordings, thus the second analysis in Section 5.3 focuses on the role of language, gestures, and eye gaze for reference resolution to objects. Subsequently, in Section 5.4 additional non-verbal channels necessary for reference resolution to objects are extracted and the interlinkage between linguistic form and non-verbal cues is analysed. Each analysis builds upon the results of the preceding analysis. The chapter will conclude with a summary and discussion in Section 5.5.

5.1 Data Collection of Multi-modal Task Descriptions

In order to investigate different aspects of task descriptions, four different tasks were developed: from the learner point of view collaboratively conducting the task, mere observing of an instructor or receiving instructions or voiced objects allowing for a more detailed analysis of information structure. The objective of the data collection is to create a corpus where a teacher explains and shows four different tasks to a learner, see Fig. 5.1 - 5.4. The developed setup is motivated by the results of the pilot study.

Letting different people explain the same tasks helps to better understand the variations of how humans naturally structure and present information. Thus, the results of the presented analyses are an important basis for what a robot would have

to deal with when it were in the learner's position. The interaction comprises of 4 tasks per instructor of about 1-2 minutes each. The reason for keeping the tasks short is to decrease the cognitive load of the teacher while explaining the tasks. Additionally, they are framed in such a way that a current robot – according to its vision and motor capabilities – would be able to perform the tasks. Moreover, the tasks were designed such that the teachers need to be explicit in their descriptions and everyday knowledge is irrelevant for understanding the teacher's instructions. The rationale for both constraints was to make the information provided in the task scenario as self-contained as possible.

5.1.1 Participants

All in all, 22 people working or studying at the Technical University Munich or the Ludwig-Maximilians-University Munich with German as their mother tongue participated in the data collection activity. In the human-human dyads, twelve male and four female teachers with an average age of 27 explained the task to a human learner. Five teachers in the HH setting interrupted their description in between and started over again (one for technical reasons and four forgot how to proceed in the middle of the task and stopped). In the human-robot dyads, three male and three female participants with an average age of 26 explained Task 3 and Task 4 to a robot learner. The instructors explaining the task to the robot were not acquainted with state-of-the-art in robotics.

5.1.2 Procedure

Recordings. The utterances of each teacher, a frontal video of the teacher, a frontal video of the learner, a video of the setting were recorded and motion was tracked for all tasks of the hands, elbows, shoulders, and head of the teachers. For the recordings three digital video cameras were used, as well as a wireless microphone worn by the teacher, a receiver, a sound mixer connected to a laptop, and Audacity[1] for audio recording. Motion was captured via Qualisys Motion Capture Systems and a Kinect sensor, see Fig. 5.1 - 5.4 for a schematic overview of the setups. In the current version of the data collection, the audio and video data of the recordings are used for analysis and annotation, whereas the motion and force data have not been analysed and annotated yet. Overall, the data collection tasks resulted in 88 recordings comprising 12 human-robot (six in Task 3 and 4 respectively) and 76 human-human dyads. In 22 recordings the descriptions are directed towards

[1] `http://audacity.sourceforge.net/`

the camera (Task 1), in 54 recordings the task descriptions are directed towards a human learner (22 in Task 2, 16 in Task 3, 16 in Task 4).

Human-Human (HH) Dyads. The first task presentation was directed towards a camera with the instruction that a person watching the video should be able to conduct the task. The second, third and fourth tasks were directed towards a human learner, who was told to carefully watch and listen to the explanations of the learner to be able to pass the information on to a new learner. In the subsequent trial, the learner became the new instructor. A calibration trial was introduced at least after every fifth trial where the experimenter functioned as an instructor to counteract the Chinese whispers effect (i.e., that the task descriptions get altered over time). The experimenter used the same wording each time. Additionally, before each task the teachers received a schematic "cheat sheet" depicting the course of action during the task to reduce their cognitive load. The aim of introducing a calibration trial and presenting the task at least after every fifth trial using the same words was also introduced to minimize variations in wording and to investigate which amount of variation is nevertheless present in the data.

Human-Robot (HR) Dyads. In the HR setting, the task was explained to the participants by the previous learner in the human-human dyads or by the experimenter. They also received a "cheat sheet" depicting the course of action. Instructors participating in the HR dyads explained the first task into the camera, the second task to a person and the third and fourth to a robot. The robot employed was a research prototype developed at the Institute of Automatic Control Engineering at the Technical University in Munich. It is of human-size height and is equipped with an omni-directional mobile platform, two anthropomorphic arms, and a pan-tilt unit on which Kinect sensors are mounted. Movement, head movements, and verbal feedback (e.g., *ja*, *ok*) were controlled by a human wizard. Empirical evidence has shown that non-verbal feedback from listeners such as eye gaze communicates understanding and is expected by human speakers (Eberhard et al., 1995). Additionally, speakers who do not get feedback from addressees take longer and make more elaborate references (Krauss & Weinheimer, 1966). Therefore we employed head-movements of the robot (so that the speaker was able to infer its eye gaze) and verbal backchannel feedback. The Kinect mounted on top of the robot (its "head") was controlled by a Wizard-of-Oz during the task descriptions and directed either towards the setting or towards the face of the instructor. Additionally, the

MARY Text-to-Speech Synthesis platform[2] was employed for giving verbal feedback during the task. For technical reasons, verbal feedback worked only for five of the six participants.

As not all instructors learned the tasks from participants, but from the experimenter, additional learners were required. Due to organisatory reasons five instructors explained the task to a "knowing" learner, who was already acquainted with the task but instructed to act as if he/she did not know the task. In this data collection, the focus is on the information transmitted by the instructor. Although "knowing" learners might react differently than naïve learners, I argue that for the research questions of this thesis it is sufficient that the instructor assumes that he/she is explaining the task to a naïve learner.

Questionnaire

In the HR dyads, the participants were additionally asked to fill in a questionnaire about their knowledge with state-of-the-art in robotics and speech synthesis as this might influence their assessment of the robot and the interaction in general. They were asked:

- whether they have worked with robots before and if yes, in what context;
- if they had the impression that there was a human or an algorithm behind the robot's navigation;
- if they had the impression that there was a human or an algorithm behind the robot's verbal feedback and head movements;
- whether they have worked with speech synthesis before;
- to rate the naturalness of the interaction with the robot on a five point Likert scale.

5.1.3 Task scenarios

In the following, the tasks are described, including (i) the course of action, (ii) how the learner and the instructor are involved in conducting the task, (iii) the items included in the task, and (iv) the reason for developing the accordant task.

Task 1. In the first task, the instructor is standing in front of a table and manipulates objects on the table. There is no learner present. The items to be manipulated are a white sheet of paper on the left side of the instructor and a plate with three wooden pieces of fruit (a banana, a strawberry and a pear) on the right side, see

[2] `http://mary.dfki.de/`

Figure 5.1. Additionally, the instructor is equipped with a second sheet of paper depicting six steps of putting the pieces of fruit on certain locations on the paper and then reordering them. The instructor first describes the initial situation and then explains in the camera how to order the pieces of fruit from the plate on the white sheet of paper. One after the other, the three pieces of fruit are put on certain locations at the paper. Subsequently, two re-ordering movements of the pieces of fruit on the paper are conducted and the locations of two pieces of fruit changed.

This task was developed with a focus on auditory perception. All objects' names are voiced in order to produce audio recordings suitable for investigating information structure (e.g., prosody, givenness, focus).

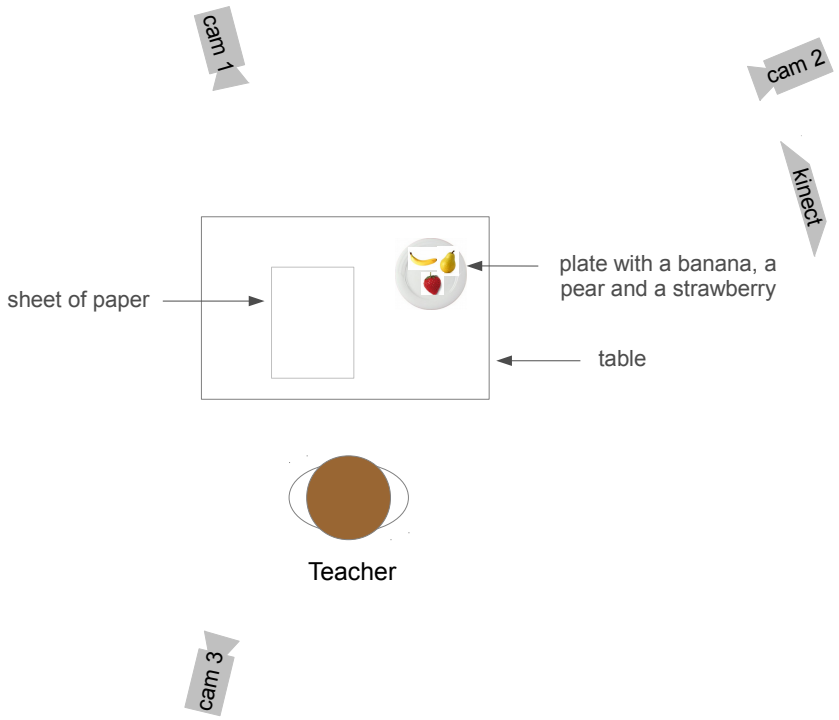

Figure 5.1: Task 1: the task setup for ordering fruits (HH: n=22).

Task 2. The goal of Task 2 is for the instructor and the learner to collaboratively move an object, standing at a table opposite of each other. On the table between the two participants, there is a board with two handles, see Figure 5.2. One handle is directed at the instructor and the other one at the learner. Both handles are marked with colours. When the task starts, the instructor asks the learner to grasp the handle at the learner's side with the left hand. The instructor grasps the handle at his/her side with the right hand. Then they lift the board and change position, i.e., they move around the table 180 degrees. Subsequently, they tilt the board 90

degrees, move along the table to the left side of the learner (i.e., the right side of the instructor), put the board down on the floor and lean it against the table.

For this task, the focus is on collaborative movement of one object. In addition to explaining and conducting the task, the instructor has to observe whether the actions of the learner are correct.

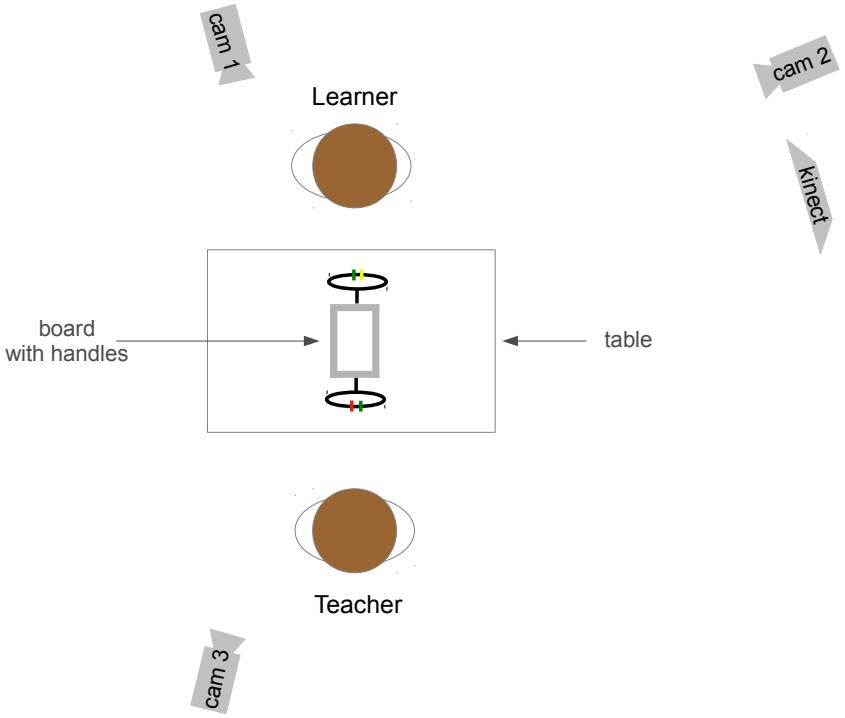

Figure 5.2: Task 2: the task setup for collaboratively moving an object (HH: n=22).

Task 3. In the third task, an instructor explains and shows to a learner how to connect two separate parts of a tube and then to mount the tube in a box with holdings. The learner stands in front of the table at the left side of the instructor (see Figure 5.3) and observes the task. Objects involved are a box with holdings placed on a table, a part of the tube already attached to the box and a loose part of the tube on an additional small table on the right side of the instructor. The loose part of the tube contains two coloured markers: a green and yellow one and a red and yellow one. First, the instructor grasps the loose part of the tube on the right side with the right hand. This part must then be connected at the green and yellow marker with the part of the tube attached to the box. The tube then must be placed between two green holdings at the green and yellow marker. Subsequently, the tube must be grasped at the red and yellow marker and put between the other pair of green holdings.

The learner is only observing while the instructor is explaining and conducting the task. Therefore the learner has less influence on the task description than in Task 2.

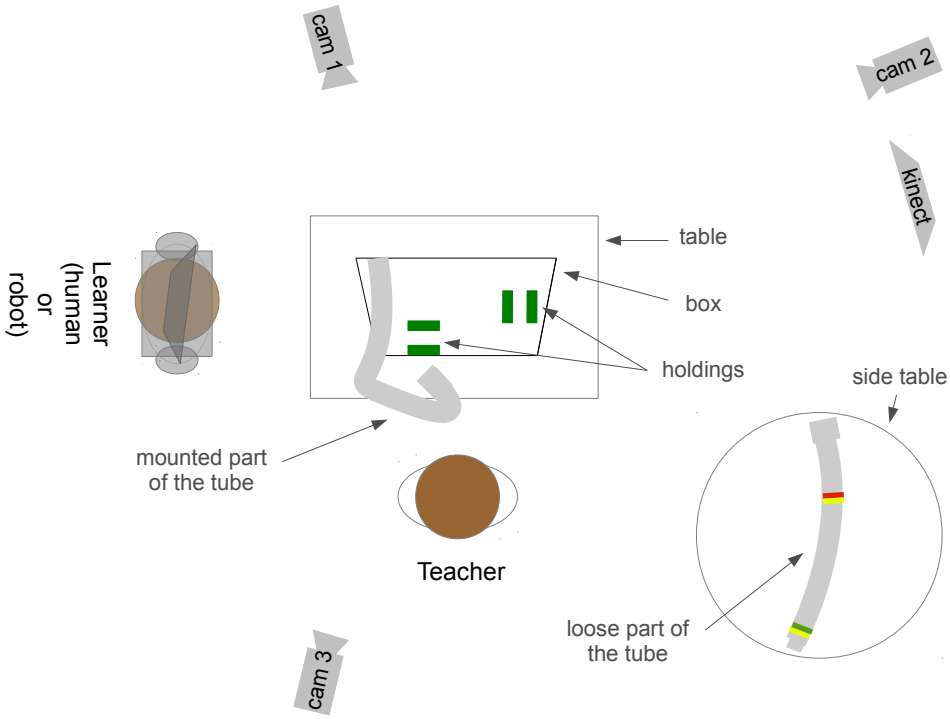

Figure 5.3: Task 3: the task setup for mounting a tube (HH: n=16; HR: n=6).

Task 4. The fourth task is a navigation task. The learner is instructed which path to take to reach a chair. In the room, there is a square table, a round table, a chair, and a small ball lying on the chair. Before the task starts, the learner is standing next to the square table, see Figure 5.4. The learner then has to pass the long side of the table, then the short side. Subsequently, the instructor asks the learner to walk around the round table towards the chair but does not say in which direction. The path on the left side and the path on the right side are equidistant. When the learner initiates to move around the table in a certain direction, the instructor corrects him/her to walk around the table in the other direction. The learner then has to look at the chair and check if there is an object located on it.

The instructor explains and the learner conducts the task. Additionally, a correction is included in the corpus.

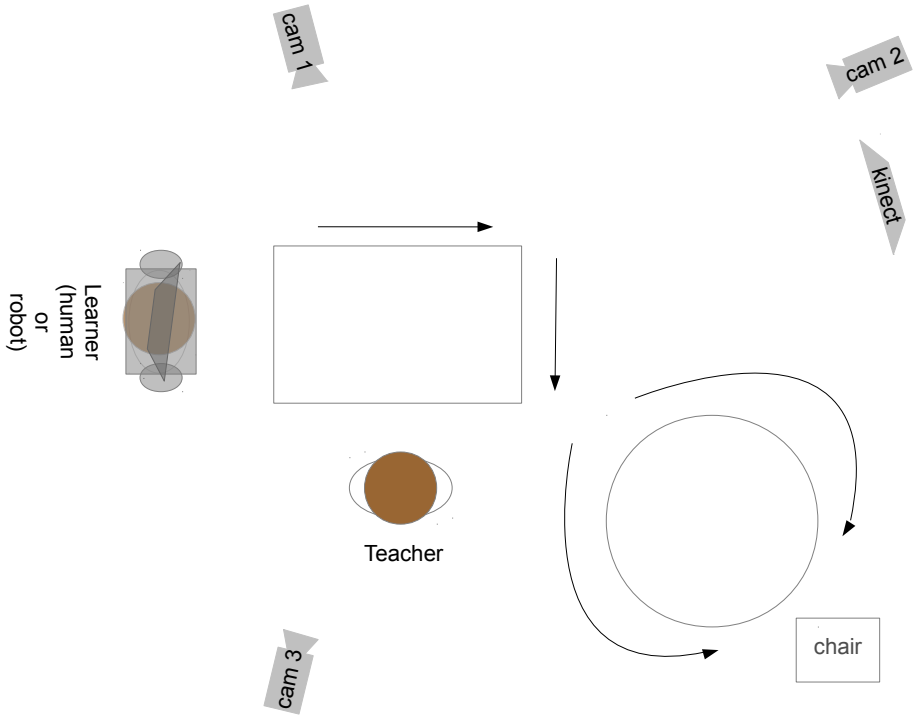

Figure 5.4: Task 4: the setup for the navigation task (HHI: n=16; HRI: n=6).

Limitations of the data

Size of the data. The size of the data is rather small. In the HH dyads, only 22 instructors in Task 1 & 2 and 16 in Task 3 & 4 explained the task to a learner. The sample size is sufficient for a qualitative analysis, and to include frequencies and percentages, but too small to employ statistical tests.

Familiarity. Not all instructor-learner dyads knew each other before. However, the task description was quite structured and participants did not talk about other topics than the task during the description. Therefore I argue that the differences in familiarity between instructor and instructee did not negatively influence the interaction.

Generalisation between tasks. Some generalisations have to be treated with caution. It also influences the task description by the instructor whether the learner is actively integrated in the task or takes only an observing role.

Participants are not balanced. Only students or people working at the university and more men than women participated (HHI: 15:7, 12:4). This is a common problem due to limited access to participants and resources.

5.1.4 Tools used for annotation

For data analysis, the audio and video data were converted into a suitable format. There is a wide range of annotation tools suitable for different kinds of data with different pros and cons, see Section 3.3.3 for a summary, as well as (Abuczki & Ghazaleh, 2013) and (Gries & Newman, 2013).

For the annotation and synchronisation of the data, the following tools were selected:

- ELAN[3] was employed for the remaining manual annotations and for synchronising audio, video and representation tiers, thus, supporting analyses across modalities (Wittenburg et al., 2006).
- Praat[4] was used for transcribing the utterances and annotating prosodic information (Boersma, 2002). The Praat tiers were then imported in Elan for further analysis.
- TreeTagger[5] was used to tag the data for part of speech (i.e., the grammatical class of words such as noun, pronoun, verb, adverb etc.) (Schmid, 1995).

In addition, Python programs were written to automatically extract temporal sequences of object references and respective cues on the different modalities. A mixture of quantitative and qualitative methods was used. For the quantitative analysis, only frequencies and percentages were used, as the data size is too small to conduct statistical tests.

5.1.5 Annotation schemes

The different layers of information annotated in the data collection are described in the following and sample annotations are presented. For all annotation tiers, a set of labels was pre-defined to denote the different objects, e.g., "loose part of the tube" (*lose Schlauch*) for the part of the tube with the green and yellow marker laying on the table at the beginning of Task 3 or "fixed part of the tube" (*fixierte Schlauch*) for the part of the tube mounted in the box.

Transcription of instructor utterances. First, the sound files with the utterances were manually transcribed, using graphemic representation, being as close as possible to the spoken utterance, i.e., keeping:

[3] `https://tla.mpi.nl/tools/tla-tools/elan/`

[4] `http://www.fon.hum.uva.nl/praat/`

[5] `http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.en.html`

- *disfluencies* such as fillers, e.g., *ähm, äh* ("ehm, eh"), false starts *ins Mitt, in die Mitte* ("in the mid, in the middle"), repetitions e.g., *dass ähm dass* ("that ehm that");
- *dialectal utterances*, e.g., *na des hebt net* for *nein, das hält nicht* ("no, this does not keep together");
- *concatenations of words*, e.g., *erklärs* (standing for *erkläre es*, "explain it");
- *elisions*, e.g., *erklär* instead of written *erkläre*.

The transcriptions were made in Praat to allow optimal temporal alignment of speech signal and transcription and have then been imported to Elan, see Figure 5.5 for an example.

Transliteration. In addition to the transcription, an extra layer of text is added where concatenations typical for spoken language are separated, elisions are recovered, etc. so that the utterances are as close to written text as possible. At this layer the spoken unit *erklärs* from the transcription layer is separated into the two words *erkläre* ("explain") and *es* ("it").

POS. The transliterated utterances were used as input to the TreeTagger Schmid (1995) and the thus resulting part-of-speech sequences were imported to Elan and manually corrected. See line 3 of Table 5.1 for the annotations on the POS-tier. The labels stem from the Stuttgart-Tübingen Tagset[6].

The example in Figure 5.1 is taken from Task 3 where the instructor attaches the end of the tube with the red and yellow marker to the left pair of green holdings. Line 1 shows the transcribed utterance *und dann wos rot-gelb is*. (The full utterance is *und dann wos rot-gelb is in die Halterung* ("and then where it red-yellow is into the holding").) Line 2 shows the transliteration where *wos* is separated into *wo* and *es*. Line 3 shows the respective parts-of speech.

1	und	dann	wos		rot-gelb	is
2	und	dann	wo	es	rot-gelb	ist
("and then where it red-yellow is")						
3	KON	ADV	PWAV	PPER	ADJD	VAFIN

Table 5.1: Sample annotation: transcription-, transliteration- and POS-tier

In addition to verbal information, non-verbal cues are annotated.

[6]http://www.ims.uni-stuttgart.de/forschung/ressourcen/-lexika/TagSets/stts-table.html

Gesture of the instructor. There exists a number of coding schemes for non-verbal behaviour, some of which are rather extensive such as the MUMIN (Allwood et al., 2007) and the BAP (Dael et al., 2012) coding schemes. The chosen coding scheme for gestures was adapted to the requirements of the data which comprises mainly object manipulation and pointing gestures. While in most gesture coding schemes, three phases of gestures "preparation", "stroke", "retraction" are annotated (see e.g., Kendon, 2004), this granularity is not needed to investigate the research questions presented in this thesis. The gesture annotation starts in the middle of the preparation phase and ends in the middle of the retraction phase. In the coding scheme pointing, iconic gestures (depicting aspects of objects, actions, etc.), beat gestures (spontaneous gestures when speaking), emblem gestures (symbolic gestures substituting words), exhibiting gestures (e.g., raising an object in order to direct the interlocutors attention on it) and poising gestures (e.g., poising with the hand above an object before grasping it) produced by the instructor are manually annotated. In addition, for (i) pointing gestures, the object, location or person the gesture is directed at is annotated, for (ii) iconic gestures, the accordant action, for (iii) emblem gestures the kind of emblem that is used (e.g., "thumbs up" for "great"), (iv) for exhibiting gestures, the object emphasised by the gesture and for (v) poising gestures also the object emphasised by the gesture.

Eye gaze of the instructor. Where (to which object, location or person in the scenario) the instructor is looking is manually annotated, as there was no eye tracker available during the recordings. Different from gestures, there is a continuous annotation of eye gaze over time, because the instructors continuously looked somewhere.

Relevant objects. On the "relevant objects"-tier the salient objects in the respective task description scene (excluding the learner/listener) were manually annotated.

For each task, a list of relevant objects was made. In Task 3 ("mounting a tube in a box with holdings"), for instance, the following objects are involved and, thus, need to be set into focus by the instructor for the learner to be able to follow the task: a loose part of the tube, a mounted part of the tube, the two parts connected to one tube, a green and yellow marker, a red and yellow marker, a pair of green holdings at the right side of the instructor, and a pair of green holdings at the left side of the instructor.

On the "relevant objects"-tier the time span a specific object is salient is marked and the time span is labelled with the respective object label. In addition to the

concrete objects involved in the task scenario, there is also a label provided for the task itself, as it is typical for the data that the instructors refer to the task itself, typically at the beginning and the end of the task description, e.g., uttering "the task is the following" (*die Aufgabe ist folgende*).

In the course of the analysis, two additional non-verbal cues were added that seemed to be valuable for reference resolution: "holding object" and "still holding object". The following tiers were only annotated for Task 3, human-human dyads.

Holding object and still holding object. "Holding object" and "still holding object" are annotated on the same tier.

> If an instructor just grasped an object and is currently holding it, it is annotated with the same tags as for relevant objects. The annotation starts as soon as the instructor's hand touches an object to hold it and ends when it is released, or when the instructor grasps a new object and still holds on to the old one.
>
> If an instructor is grasping several objects at the same time, they are separated by ' / ' in the annotation tier.
>
> If an instructor is holding an object, grasping a new one and still holding that previous object, the new object is written first, the old one is also annotated in brackets.

For example the instructor grasps the loose part of the tube at the green and yellow marker, "lose Schlauch / grün-gelbe Markierung" is annotated. As soon as the instructor touches with his/her other hand the mounted part of the tube in order to grasp it, the annotation changes to "fixierte Schlauch (lose Schlauch / grün-gelbe Markierung)" and the loose part of the tube as well as the green and yellow marker are thus annotated as "still holding object" and the mounted part of the tube as "holding object".

The salience of an object is identified either by the occurrence of a linguistic reference in the instructor's speech, by the instructor's gaze behaviour, specific communicative gestures such as deictic gestures, using fingers for counting, raising the index finger when talking about something important, or whether the instructor is holding or still holding an object. Linguistic indicators are, for instance, full or elliptic noun phrases, e.g., "the tube" (*den Schlauch*), "tube" (*Schlauch*), pronouns, e.g., "it" (*er*), "the" (*der* for "the tube" (*der Schlauch*), determiners combined with spatial indexicals, e.g., "the one here" (*den hier*), spatial indexicals, e.g., "here",

Figure 5.5: Elan Screenshot. An instructor is mounting a tube in a schematic motor block.

"there" (*hier, da*), adjectives, e.g., "red-yellow" (*rot-gelb*) for the red and yellow marker attached to the tube. For examples of salient objects, see Tables 5.2 and 5.3, line 4. In the first example (Table 5.2), linguistic indicators for the salient object "end of tube with red-yellow marker" are the spatial indexicals *wo*, the personal pronoun *es* and the adjective *rot-gelb*. In Table 5.3, the salient object is "the green holdings to the left of the instructor" co-occurring with the noun phrase *die Halterung*.

Examples for linguistic indicators that make the task itself salient are "the task is about" (*also hier geht es darum*) which is typically used at the beginning of a task presentation, and "this was it" (*das wars*) to indicate that the task presentation is now finished.

1	und	dann	wos		rot-gelb	is
2	und	dann	wo	es	rot-gelb	ist
('and then where it red-yellow is')						
3	KON	ADV	PWAV	PPER	ADJD	VAFIN
4			red yellow marker			

Table 5.2: Sample annotation: transcription-, transliteration-, POS- and "salient object"-tier

1	in	die	Halterung
2	in	die	Halterung
('into the holding')			
3	APPR	ART	NN
4		left-side green holdings	

Table 5.3: Sample annotation: transcription-, transliteration-, POS- and "salient object"-tier

Prosodic information was annotated according to the DIMA annotation guidelines (Kügler et al., 2015). This annotation schema has been chosen because it (i) represents a consensus system for prosodic annotation of German, (ii) aims at compatibility of annotations and thus fosters the exchange of annotated data sets, and (iii) allows for independent annotation of phrase boundaries, prominence levels and tones. As regards the data collection presented in this thesis, phrase boundaries and prominence levels are annotated:

Phrase boundary. Phrase boundaries were annotated and differentiated based on auditory-phonetic criteria such as pauses, final lengthening, tonal movement, pitch reset. Weak (-) and strong (%) boundaries were distinguished, and constitute a hierarchical structure, whereby a phrase with weak boundaries is dominated by a phrase with strong boundaries.

Prominence level. Subsequently, prominent syllables were annotated with levels of perceived prominence. DIMA proposes three levels of prominence:

Prominence level 1 (weak prominence) refers to metrical strength and tonal events such as rhythmic accents, phrase accents, post-lexical stress, etc.
Prominence level 2 (strong prominence) refers to pitch accent.
Prominence level 3 (emphasis, extra strong prominence) refers to attitudinal emphasis beyond the prominence of pitch accents.

For an exhaustive presentation of the different tiers of prosodic DIMA annotation (see Kügler et al., 2015). Praat has been used for making the prosodic annotations and an example from the data is shown in Figure 5.6.

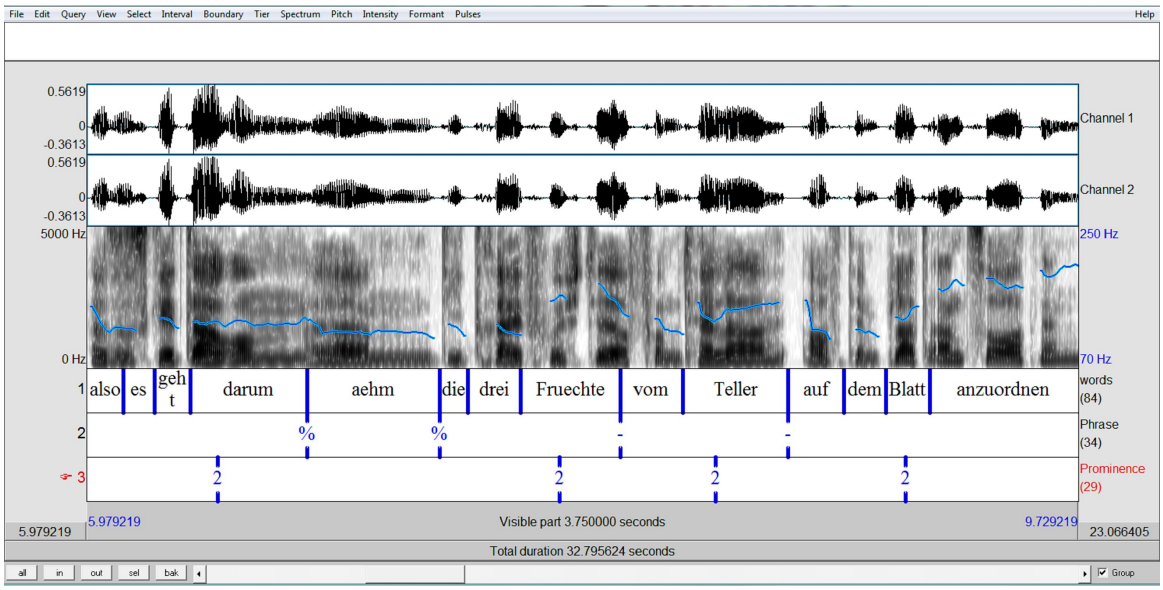

Figure 5.6: Sample annotation in Praat: phrase boundaries and prominence levels. % indicates strong phrase boundary, - weak boundary, and 1 and 2 stand for prominence levels 1 and 2. The annotation example is taken from Task 1.

Due to the qualitative approach of this thesis, the data was collected in a way allowing the investigation of different aspect of situated task description. Depending on the results of the first analyses and the research questions of this thesis, only a subset of the data collection is further analysed. However, parts of the rich data material which are not considered in this thesis (such as motion data and information structure) can serve as an empirical basis for potential future work.

5.1.6 Inter-coder agreement

The whole data set has been annotated by the author of this thesis. In the first explorative analysis, all four tasks are investigated. However, in the second and

the third analysis presented in this thesis, there is a focus on human-human object references and thus on Task 2 and Task 3. These two tasks include different objects and the learner is physically present. This subset of 22 human-human dyads in Task 2 and 16 human-human dyads in Task 3 has been additionally annotated by a linguist not involved in the work of this thesis. An important prerequisite for shedding light on the importance and interplay of different modalities is to determine which object is actually intended by the participant with a certain referring expression. For example, participants might utter "You have to grasp this thing" or "You have to grasp the green and yellow marker". Both "the thing" and "the green and yellow marker" are object references which need to be resolved to a physical object. Thus, on the object tier, for all object references the annotator had to label the intended object (based on her competence) on the "relevant objects"-tier. Cohen's kappa was computed to measure inter-rater agreement for the relevant object. With 0.918 the kappa coefficient agreement between annotators is high, showing that humans are rather consistent in interpreting multi-modal references to objects.

5.2 Explorative Analysis of Inter- and Intra-speaker Variation

The first approach to the data was a general, explorative analysis of variation between and within task descriptions. The viewpoint taken was that of a robot being exposed to human multi-modal task descriptions. Aim of this first analysis was to identify (i) general patterns that could be used by the robot to extract information with regards to the task, as well as (ii) potential challenges for robots. The analysis is based on the results of the pilot study.

5.2.1 Research questions

In particular, the following research questions were addressed:

RQ1 On which channels is relevant information transmitted?

RQ2 Which phenomena or general patterns occur during task descriptions and what is their impact on comprehension?

RQ3 What is the inter- and intra-speaker variability in conveying respective information?

RQ4 What are the differences in how a task is transmitted between human-human and human-robot dyads?

5.2.2 Results

The learner had the assignment to listen carefully and forward the information to a new learner. Even though the tasks were quite simple, there was considerable variation in how instructors structured and presented the task. The multi-modal qualitative analysis of the data revealed:

(i) characteristics of spoken language, such as insertions, interruptions, hesitations etc.
(ii) variation in wording regarding objects and actions (i.e., different nouns or adjectives plus nouns for objects and different verbs for actions), as well as omissions of lexical referents,
(iii) temporal structuring of the task by verbal means,
(iv) variation in the perspective taken by the instructors, i.e., not literally interpretable personal pronouns, and
(v) patterns of use of verbal references and/or communicative gestures for directing the learner's attention.

Characteristics of spoken language. In all four tasks in the human-human and the human-robot dyads, several characteristics of spoken language occurred. Except for one instructor explaining Task 1 towards the camera and one instructor explaining Task 4 to a robot learner, each task description of each instructor contained disfluencies of spoken language. The disfluencies include (i) non-lexical fillers (e.g., "äh"), (ii) lexical fillers (e.g., "like this" (*so*)), (iii) abandoned utterances (e.g., "this is somehow" (*das ist irgendwie*)), (iv) repairs or corrections (e.g., "next to well on the right side of the banana" (*neben die also rechte neben die Banane*)), (v) contractions (e.g., "that's it" (*das wars*, "das war es")), (vi) repetitions (e.g., "table table" (*Tisches Tisches*)), (vii) omissions (e.g., "next table" for "next to the table" (*neben Tisch*)), and (viii) dialect (e.g., "no this is not working" (*na des hebt net* for 'nein das hält nicht')).

Disfluencies might hinder the correct interpretation of utterances, especially when disconnected from visual information.

In the following, the variation in wording, temporal structuring of the task by verbal means, gestures, and eye gaze of the instructors are discussed for each task, divided in human-human and human-robot dyads.

Human-Human Dyads

22 participants instructed a human learner in Task 1 and 2, and 16 in Task 3 and 4.

Task 1

On average, the task duration was 56 seconds (22 sec - 2 min 3 sec). The 22 instructors were asked to carry out the task and explain each step in detail towards the camera. In the following, prevalent phenomena are presented and discussed.

Variation in wording. Relevant entities and actions occurring in the task are: a plate, three wooden pieces of fruit (a banana, a pear, and a strawberry), a piece of paper, grasping the pieces of fruit, putting them on the paper, and re-ordering them.

Objects mentioned by all participants were the banana, the strawberry, the pear, and the sheet of paper. The action most important for the task and the only one mentioned by all participants was putting the pieces of fruit on the paper. While the wordings for the different pieces of fruit were consistent, they varied for the other objects and up to eight different noun phases were used for the plate, see Table 5.4. However, for verbs the variation was even higher: up to 13 different verbs were used for the action of putting the pieces of fruit on the sheet of paper.

Table 5.4: Task 1 - Summary of the wording (n=22). Concepts mentioned by at least 5 participants are listed. ∅ indicates the number of participants omitting a NP for referring to the accordant object.

Object/ Action	**Wording**
Pieces of fruit	'pieces of fruit' (*Früchte*) (4), (*Obst*) (5), (*Obststücke*) (1), 'wooden fruits' (*Holzfrüchte*) (1), 'fruit variety' (*Obstsorte*) (1), 'the whole' (*das Ganze*) (1), ∅ (10)
Banana	'banana' (*Banane*) (21), 'yellow banana' (*gelbe Banane*) (2)
Strawberry	'strawberry' (*Erdbeere*) (21), 'red strawberry' (*rote Erdbeere*) (2)
Pear	'pear' (*Birne*) (22), 'green pear' (*grüne Birne*) (1)
Plate	'plate' (*Teller*) (22), 'white plate' (*weißer Teller*) (1), ∅ (4)
Paper	'sheet' (*Blatt*) (19), 'sheet of paper' (*Blatt Papier*) (5), 'white sheet of paper' (*weißes Blatt Papier*) (2), 'sheet of paper of the size DINA4' (*Blatt Papier der Größe DINA4*) (1), 'empty sheet' (*leeres Blatt*) (1), 'DINA4 sheet' (*DINA4 Blatt*) (1), 'white sheet' (*weißes Blatt*) (1), 'empty white sheet' (*leeres weißes Blatt*) (1)

Table 5.4: (continued)

Object/ Action	**Wording**
Take	'take' (*nehmen*) (19), ∅ (3)
Put	'put' (*legen*) (19), (*drauflegen*) (1), (*packen*) (1), 'place' (*platzieren*) (5), 'come' (*kommen*) (5), 'do' (*tun*) (2), 'put down' (*absetzen*) (2), 'order' (*anordnen*) (1), (*order*) (1), 'arrange' (*arrangieren*) (1), 'move upwards' (*nach oben rücken*) (1), 'jump over' (*überspringen*) (1), 'change order' (*umlegen*) (1)
Lie	'lie' (*liegen*) (8), (*zu liegen kommen*) (1), 'have' (*haben*) (8), 'arranged' (*befinden*) (3), 'lie upon' (*darauf liegen*) (1), 'stand' (*stehen*) (1), 'be' (*sein*), ∅ (6)
Centre of the page	'middle' (*Mitte*) (19), 'centre' (*Zentrum*) (2), ∅ (2)
Step	'step' (*Schritt*) (6), 'order' (*Reihenfolge*) (1), 'time' (*Mal*) (1), ∅ (15)
Corner of the page	'corner' (*Ecke*) (3), 'border' (*Blattrand*) (1), 'end' (*Ende*) (1), ∅ (17)
Order	'order' (*bestimmte vorgegebene Reihenfolge*) (1), (*Veränderung der Reihenfolge*) (1), (*Reihenfolge*) (1), 'arrangement' (*Arrangement*) (1), 'row' (*Reihe*) (1), ∅ (18)
Change	'move' (*verschieben*) (3), 'change' (*vertauschen*) (2), (*verändern*) (1), 'correct' (*korrigieren*) (1), 'change location' (*Plätze tauschen*) (1), ∅ (16)
Finished	'finished' (*fertig*) (4), (*damit sind wir fertig*) (1), (*damit wären wir fertig*) (1), 'that's it' (*das wars*) (3), 'come to the end' (*am Ende angekommen*) (1), ∅ (12)

Time Markers. All instructors used time markers to structure the task in smaller sequences, e.g., "now" (*jetzt*), "then" (*dann*), "subsequently" (*danach*), "the next step" (*als nächstes*). Four out of 22 instructors conducted a posture shift (moving either towards of away from the setting to signal its start or ending) and three nodded towards the experimenter when they had finished the task.

Initiating moves "so" (*also*), "ok", "good" (*gut*) were conducted by 14 instructors. 13 verbally communicated when the task was done: "that's it" (*das wars*), "done" (*fertig*), "good" (*gut*), "then we have reached the end" (*dann sind wir am Ende angekommen*).

Instructors' perspective. The perspective preferred by half of the instructors (11) was 1^{st} person singular (*ich*), two used 1^{st} person plural (*wir*), one 3^{rd} person singular (*man*), and one 2^{nd} person singular (imperative), see Table 5.7. The other seven instructors changed the perspective during the task description: three changed from 3^{rd} person singular to 1^{st} person singular indicative voice and two changed back to 3^{rd} person singular, two used 3^{rd} and 1^{st} person plural, and two changed back and forth between 1^{st} person singular and 1^{st} person plural.

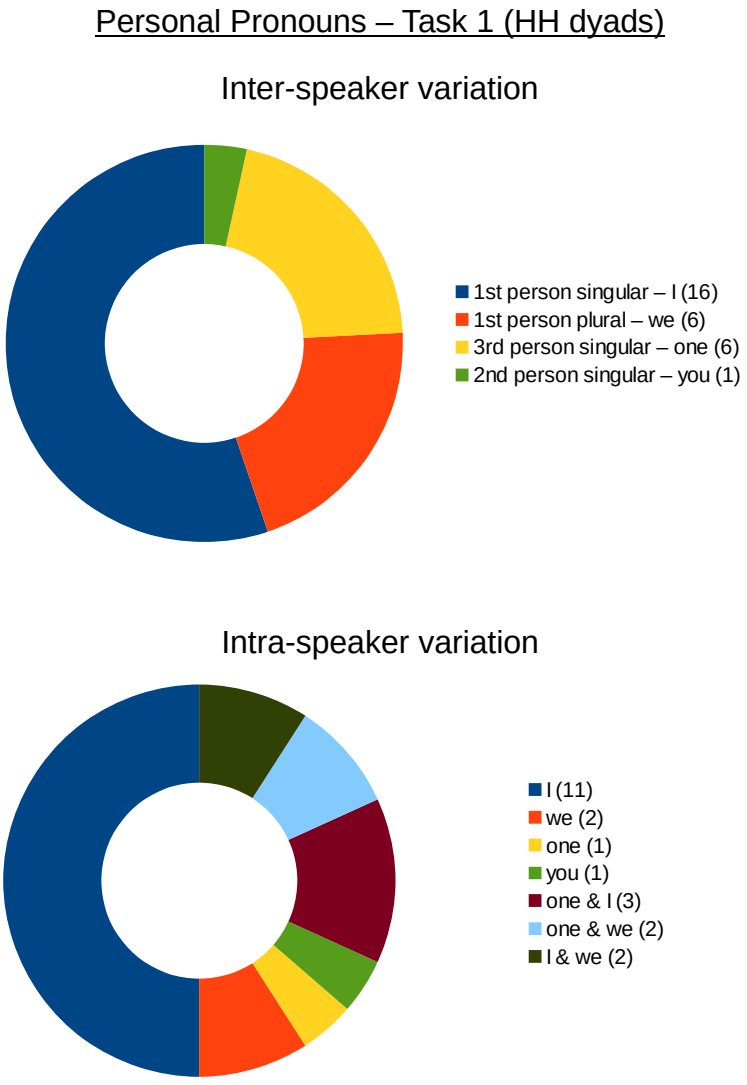

Figure 5.7: Task 1: Personal pronouns uttered between and within tasks (n=22).

Verbal and gestural references to the scene. 14 out of 22 instructor employed verbal references to the scene or gestures to transmit information during the task description. 10 instructors uttered verbal references which had to be resolved via

the visual scene, e.g., "here" (*hier*). Amongst these, 5 instructors referred to some of the entities or locations using gestures. Nevertheless, seven referred to an object or a location, e.g., via "here", "there" (*hier*, *da*) without resolving it via language or gestures. In these cases, eye gaze was directed at the intended object or location and thus a potential cue to resolve ambiguities. Half of the instructors used supplementary gestures – either pointing gestures or exhibiting gestures – in addition to the verbal descriptions of the task.

Indexicals (verbal references to the visual scene) were uttered by all instructors, referring to (i) locations, e.g., "here", "there" (*hier*, *da*), (ii) the manner in which the task was conducted, e.g., "like this" (*so*), or (iii) objects, where visual information is needed to resolve the ambiguities, e.g., "this part" (*dieses Teilstück*), "this green thing" (*dieses grüne Ding*).

Eye gaze. There was no learner present, thus the instructor could not direct the eye gaze towards the learner, but only towards the experimenter or towards the camera. Two instructors frequently looked towards the camera (both four times) and two directed their gaze towards a non-existing audience.

Task 2 The average task duration was 36 seconds (17 sec - 1 min), 22 instructors participated. The instructor and the learner had to collaboratively move an object.

Variation in wording. In the task, the instructor had to use the right hand and the learner the left hand. While the "right hand" was uttered by half of the instructors, the left hand used by the learner was uttered by all. The handle of the board to be grasped, one by the learner and one by the instructor, was uttered by 13 instructors using 7 different wordings (e.g., "handle" (*Griff*, *Henkel*, *Hantel*)). The board was the object to be collaboratively moved during the task. Eleven different wordings were uttered for the board and two instructors did not utter a verbal reference for this object at all. Regarding the nine different nouns by the other 20 instructors, a bandwidth from very specific to very unspecific words was used. While some participants uttered more unspecific words for the board, such as the pronouns "the/it" (*das/es*), "the whole" (*das Ganze*), "the thing" (*das Ding*), others used more concrete referents, e.g., "board" (*Brett* or *Balken*). Four instructors used pronouns without verbal antecedents when talking about the object to be manipulated. In these cases, the reference could solely be resolved via visual antecedents.

For the table (*Tisch*), the nouns did not vary between participants. Regarding lifting the board, 10 instructors uttered four different verbs and for tilting the board, 14 instructors uttered eight different verbs. The number of inter-speaker

variation even increased for the actions of walking around the table and leaning the board against the table. All instructors verbally described these two actions: eleven different verbs were used for walking around the table, e.g., "walk around" (*herumgehen)*, "walk" (*gehen*), "turn" (*drehen*), and 13 for leaning the board against the table, e.g., "put down" (*abstellen*), "lean against" (*anlehnen*), "lean" (*lehnen*), up to two by each instructor, , see Table 5.5.

Summing up, the objects mentioned by all instructors were the left hand and the object to be manipulated – the board. Eleven different wordings were uttered for the board and four instructors used a pronoun when they first mentioned it. For actions, walking around the table and placing the board next to table were uttered by all. Again, a broad variety of wordings occurred for these two actions.

Table 5.5: Task 2 - Summary of the wording (n=22). Concepts mentioned by at least 5 participants are listed. ∅ indicates the number of participants omitting a NP for referring to the accordant object.

Object/ Action	**Wording**
Right hand	'right hand' (*rechte Hand*) (11), ∅ (11)
Left hand	'left hand' (*linke Hand*) (20), 'your left' (*deine Linke*) (1), 'hand' (*Hand*) (1)
Handle	'handle' (*Griff*) (5), (*Henkel*) (2), (*gelb-grüner Henkel*) (1), (*Hantel*) (1), 'thing' (*Teil*) (1), 'green and yellow boarder' (*gruen-gelbe Umrandung*) (1), 'lever with the yellow and the green colour' (*Hebel mit der gelben und der grünen Farbe*) (1), ∅ (9)
Board	'board' (*Brett*) (9), 'thing' (*Ding*) (3), 'object' (*Objekt*) (2), 'beam' (*Balken*) (2), 'thing' (*Teil*) (3), 'item' (*Gegenstand*) (1), 'arrangement' (*Anordnung*) (1), 'it' (*das*) (3), 'the whole' (*das Ganze*) (2), 'device' (*Gerät*) (1)
Table	'table' (*Tisch*) (21), 'tabletop' (*Tischplatte*) (1), ∅ (1)
Take	'take' (*nehmen*) (13), 'grasp' (*greifen*) (6), (*anfassen*) (5), 'put hand on' (*Hand drantun*) (1), 'take hand' (*Hand nehmen*) (1), 'put hand' (*Hand legen*) (1), ∅ (1)
Raise	'raise' (*anheben*) (4), (*hochheben*) (3), (*heben*) (1), (*hochnehmen*) (1), 'take' (*nehmen*) (1), ∅ (12)

Table 5.5: (continued)

Object/ Action	**Wording**
Overturn	'assemble' (*aufstellen*) (5), (*querstellen*) (1), (*aufrecht hinstellen*) (1), (*senkrecht stellen*) (1), 'tilt' (*kippen*) (3), (*umklappen*) (1), 'turn' (*drehen*) (1), (*hochkant drehen*) (1), ∅ (8)
Walk	'walk around' (*herumgehen*) (8), (*gehen*) (6), (*um den Tisch gehen*) (1), 'turn' (*drehen*) (3), (*um den Tisch drehen*) (2), 'turn around' (*herumbewegen*) (2), 'move' (*bewegen*) (1), 'change position' (*Positionen tauschen*) (1), 'run around' (*herumlaufen*) (1), 'run' (*laufen*) (1), 'change place' (*Plätze tauschen*) (1), 'walk over' (*herübergehen*) (1), 'turn around' (*herumdrehen*) (2)
Place	'place' (*abstellen*) (11), 'lean' (*anlehnen*) (4), (*hinlehnen*) (1), (*lehnen*) (2), (*dagenenlehnen*) (1), (*ranlehnen*) (1), 'put' (*stellen*) (4), (*hinlegen*) (2), (*absetzen*) (1), (*legen*) (1), (*ablegen*) (1), (*hinstellen*) (1), 'so that its position is straight' (*dass es aufrecht steht*) (1)
Side	'side' (*Seite*) (11), 'left end of the table' (*linke Tischende*) (1), 'right edge of the table' (*rechte Tischkante*) (1), ∅ (10)
Degrees	'degrees' (*Grad*) (15), ∅ (7)
Green and yellow marker	'yellow and green marker' (*gelb-gruene Markierung*) (1), 'marker the yellow and green one' (*Markierung an die gelb-gruene*) (1), 'marker' (*Markierung*) (1), 'green and yellow marker' (*gruen-gelbe Markierung*) (1), 'marker of the yellow and green one' (*Markierung von dem gelb-gruenen*) (1), ∅ (17)
Clockwise direction	'clockwise direction' (*Uhrzeigersinn*) (7), ∅ (15)
Move	'move' (*bewegen*) (3), 'turn' (*drehen*) (3), (*herumdrehen*) (2), 'carry' (*tragen*) (2), ∅ (13)
Floor	'floor' (*Boden*) (9), ∅ (13)
Task	'task' (*Aufgabe*) (5), 'it is about' (*es geht darum*) (1), 'goal' (*Ziel*) (1), ∅ (15)
Finished	'that's it' (*das wars*) (7), 'finished' (*fertig*) (2), ∅ (13)

Time Markers. One instructor did not use any verbal time markers at all. All others used up to 8 time markers during the task description, such as "now" (*jetzt*), "subsequently" (*dann*, *anschließend*). Only one instructor took a step back in order

to signal that the task was finished. Five instructors nodded towards their learner when the task was done.

Instructors' perspective. The perspective taken by the instructor again varied from instructor to instructor as well as within one task description, see Figure 5.8. While the perspective taken in the first task did not reflect who actually carried out the task, it is reflected in the second task. All participants used 2^{nd} person singular (imperative) and 1^{st} person plural, 15 also included 1^{st} person singular, mainly when the instructor himself/herself had to grab the handle. These results suggest that in collaborative tasks the perspective taken by the person explaining is constrained to who actually has to carry out the task.

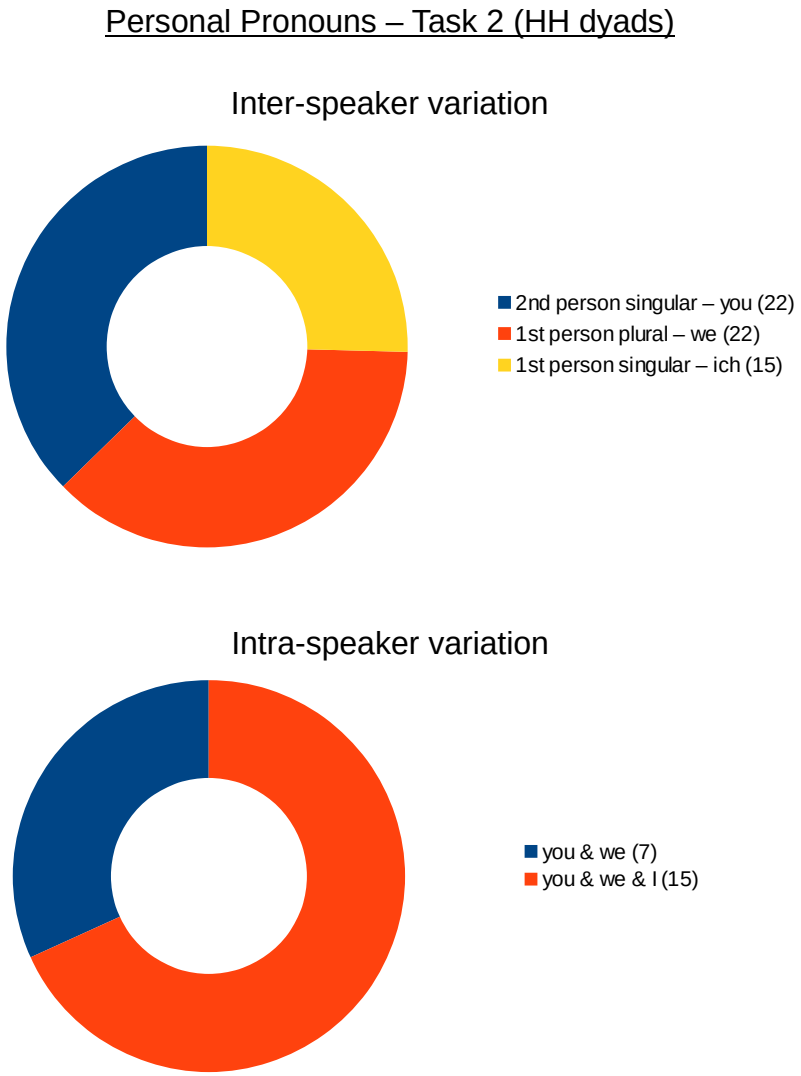

Figure 5.8: Task 2: Personal pronouns uttered between and within tasks (n=22).

Verbal and gestural references to the scene. All but one instructor used verbal references to the scene mostly in conjunction with gestures (pointing gestures, iconic gestures and beckoning the learner over). Pointing gestures were either conducted with the hand or the head.

Eye gaze. One instructor never looked at the learner during the whole interaction. All other instructors directed their eye gaze at objects relevant for the task and at the learner 4.6 times on average during the interaction and up to 11 times.

Task 3 16 instructors showed and explained Task 3 to a human learner. On average, the task duration was 41 seconds (18 sec - 1 min 48 sec).

Variation in wording. Objects relevant for the task are the loose part of the tube, the mounted part of the tube, and the two parts connected, the pair of green holdings on the right side of the learner and the pair of green holdings on the left side of the learner, the green and yellow marker, and the red and yellow marker, as well as the right hand of the instructor. Relevant actions are grasping the tube, connecting the parts of the tube, and putting the tube between the green holdings at the two markers.

In Task 3, the largest number of individual objects is involved, hence, this task is the most interesting one with regard to object references. The analysis of variation in wording shows extensive lexical variation and omitted verbal references for objects and actions. Considering what is visually perceived and what is uttered reveals how differently the same objects and actions are referred to in the utterances. Additionally, the unspoken needs to be grounded in the scene, e.g., the first pair of green holdings were not mentioned by five instructors, and the verbal expressions for the spatial perspectives varied, e.g., one instructor named the pair of green holdings on the left side of the instructor "the right one"(*das Rechte*), and another instructor named the same holdings "the left side" (*die linke Seite*).

Only the tube and the action of placing the tube were mentioned by all instructors. The frequency of mentions of an object or an action in everyday life may in part influence the variation in wording. The green barriers are less prototypical than the tube and they cause more variation in wording. Still, 9 different noun phrases for the tube were used. For actions, the possibility in German to prefix verbs with prepositions may in part explain the lexical variability for putting the tube between the green holdings, e.g., "put through" (*durchlegen*), "insert" (*hineinlegen*). Similar to Task 1 the "placing" action is more relevant for achieving the task than for example grasping the tube. However, the degree of precision in which the uttered verbs

described the action varied, e.g., "thread" (*durchfädeln*), and some expressions were more specific than others, e.g., "insert" (*hineinlegen*) versus "that it goes inside" (*dass es hineingeht*) or "green and yellow marker" (*gruen-gelbe Markierung*) versus "this side" (*diese Seite*). Also intra-speaker variation played a role in wording, e.g., one instructor used two different words for denoting the "holdings" (*Halterung*) and "barrier" (*Hindernis*), another instructor uttered four different verbs for putting the tube between the green holdings: "insert" (*hinein kommen*), "gets" (*kommen*), "put through" (*durchlegen*), "put" (*legen*), see Table 5.6.

Table 5.6: Task 3 - Summary of the wording in human-human dyads (n=16). Concepts mentioned by at least 5 participants are listed. ∅ indicates the number of participants omitting a NP for referring to the accordant object.

Object/ Action	**Wording**
Tube	'tube' (*Schlauch*) (14), 'pipe' (*Rohr*) (3), 'loose pipe' (*lose Rohr*) (1), 'the whole' (*das Ganze*) (2), 'the part/thing' (*das Teil*) (1), 'the end-piece' (*Endstück*) (1), 'the connected tube' (*verbundene Schlauch*) (1), 'the appendant parts' (*zugehörige Teile*) (1), 'the part of the tube' (*das Teil von dem Schlauch*) (1)
Right hand	'right hand' (*rechte Hand*) (8), ∅ (8)
Green and yellow marker	'green and yellow marker' (*grün-gelbe Markierung*) (5), 'yellow and green marker' (*gelb-grüne Markierung*) (3), 'green and yellow end' (*grün-gelbe Ende* (3), 'marker' (*Markierung*) (2), 'end where the green and yellow is attached' (*Ende wo das Grüne und das Gelbe dran ist*) (1), 'end with the yellow and green marker' (*Ende mit der gelb-grünen Markierung*) (1), 'yellow and green connection' (*gelb-grüne Verbindung*) (1), 'green and yellow part' (*grün-gelbe Teil*) (1), 'green and yellow section' (*grün-gelbe Abschnitt*) (1), 'this side' (*diese Seite*) (1), 'the green part/thing' (*das grüne Teil*) (1), ∅ (1)
Mounted tube	'tube' (*Schlauch*) (4), 'pipe' (*Rohr*) (1), 'segment of the tube' (*Teilstück des Schlauches*) (1), 'second tube' (*zweite Schlauch*) (1), ∅ (9)

Table 5.6: (continued)

Object/ Action	**Wording**
First green holdings	'mounting' (*Befestigung*) (1), 'this side' (*diese Seite*) (1), 'holding' (*Halterung*) (1), 'first holding' (*erste Halter*) (1), (*erste Halterung*) (1), 'first barrier' (*erste Hindernis*) (1), 'right first holding' (*rechte erste Halterung*) (1), 'green thing' (*grüne Ding*) (1), 'these two blocks' (*diese beiden Klötze*) (1), 'right green marker' (*rechte grüne Markierung*) (1), 'right channel' (*rechte Kanal*) (1), 'appliance' (*Vorrichtung*) (1), ∅ (5)
Second green holdings	'second holdings' (*zweite Halterung*) (3), (*zweite Halter*) (1), 'other green holdings' (*andere grüne Halterung*) (1), 'holdings' (*Halterung*) (1), 'other channel' (*andere Kanal*) (1), 'other appliance' (*andere Vorrichtung*) (1), 'these two' (*diese Beiden*) (1), 'side' (*Seite*) (1), 'left side' (*linke Seite*) (1), 'second green thing' (*zweite grüne Ding*) (1), 'second barrier' (*zweite Hindernis*) (1), ∅ (4)
Yellow and red marker	'red and yellow marker' (*rot-gelbe Markierung*) (5), 'the yellow and red (section/part)' (*der gelb-rote (Abschnitt/Teil)*) (2), 'yellow and red marker' (*gelb-rote Markierung*) (1), 'marker, the red and yellow one' (*Markierung, die rot-gelbe*) (1), 'red marker' (*rote Markierung*) (1), 'where it is yellow and red' (*wo es gelb-rot ist*) (1), 'this end' (*dieses Ende*) (1), 'the red one' (*das Rote*) (1), ∅ (4)
Take	'take' (*nehmen*) (11), 'work' (*arbeiten*) (1), 'grasp' (*greifen*) (1), ∅ (3)
Assemble	'assemble' (*hineinstecken*) (10), (*zusammenstecken*) (2), (*anstecken*) (1), 'assembled' (*drinnen stecken*) (1), 'connect' (*verbinden*) (4), 'combine' (*kombinieren*) (1), ∅ (1)
Put	'put through' (*durchlegen*) (4), (*durchführen*) (2), 'insert' (*hineinlegen*) (3), (*zum Liegen kommen*) (2), (*einführen*) (1), (*einlegen*) (1), (*hinein kommen*) (1), (*hineintun*) (1), (*stecken*) (1), 'install' (*verlegen*) (2), 'put' (*legen*) (2), 'that it goes inside' (*dass es hineingeht*) (1), 'thread' (*durchfädeln*) (1), 'put around' (*herumlegen*) (1), 'gets' (*kommen*) (1), 'mount' (*montieren*) (1), 'push inside' (*hineindrücken*) (1), 'clamp inside' (*hineinklemmen*) (1)

Table 5.6: (continued)

Object/ Action	**Wording**
Stay assembled	'hold' (*halten*) (3), 'be fixed' (*fest sein*) (1), 'stay assembled' (*festhalten*) (1), (*zu bleiben*) (1), (*zusammenhalten*) (1), (*schön fest bleiben*) (1), 'falls apart soon' (*fällt gleich raus*) (1), ∅ (8)
Task	'task' (*Aufgabe*) (6), 'here it is about' (*hier geht es darum*) (1), ∅ (10)
Finished	'that's it' (*das wars*) (3), 'it is finished' (*wär das erledigt*) (2), 'finished' (*fertig*) (1), 'at the end it looks like this' (*dann schaut das so aus zum Schluss*) (1), 'that was the task' (*das war die Aufgabe*) (1), 'when it stays stable the task is achieved' (*wenn das stabil bleibt ist die Aufgabe erfüllt*) (1), ∅ (7)

Time Markers. Eight out of 16 instructors told their respective learners when the task started, e.g., "So, the task is the following" [...] (*So die Aufgabe ist die folgende* [...]), "Ok here it is about" [...] (*Ok also hier geht es darum* [...]). Eleven instructors verbally signalled their learners when the task was done, e.g., "that was it" (*das wars*), "done" (*fertig*). During the task description, all instructors used verbal time markers to structure the task, e.g., "then" (*dann*), "subsequently" (*anschließend*), "now" (*jetzt*). All but two (14) instructors conducted a finalising action to demonstrate that the task was done by taking a step back when they had finished the task. One instructor alternatively made a posture shift, another one stopped in the middle when moving backwards because the connected tube fell apart.

Instructors' perspective. Also in Task 3 the instructors' expression of perspective varied in person and voice between and within speakers, see Figure 5.9. Two speakers used 1^{st} and three 2^{nd} person singular, three 1^{st} person plural and one the indefinite pronoun. One instructor used 2^{nd} person singular. The other six instructors changed the grammatical person during their explanation. Three started with either 2^{nd} person singular or 1^{st} person plural and then immediately corrected themselves: "you grasp I grasp" (*du greifst also ich greife*), "we I take" (*wir ich nehme*), "that we that one" (*dass wir dass man*). The other three varied between 1^{st} person singular and passive voice "now the tube is lead here through this marker" (*jetzt wird der Schlauch hier durch diese Markierung durchgeführt*), between 1^{st} and 2^{nd} person singular and between 1^{st} person plural, 1^{st} person singular, indefinite pronoun, and

passive voice.

Three learners interpreted the 2^{nd} person singular perspective of the instructor wrongly and initiated movements to become active in conducting the task. But they were either corrected by the instructor or the instructors ignored them and they then realised that they only had to observe.

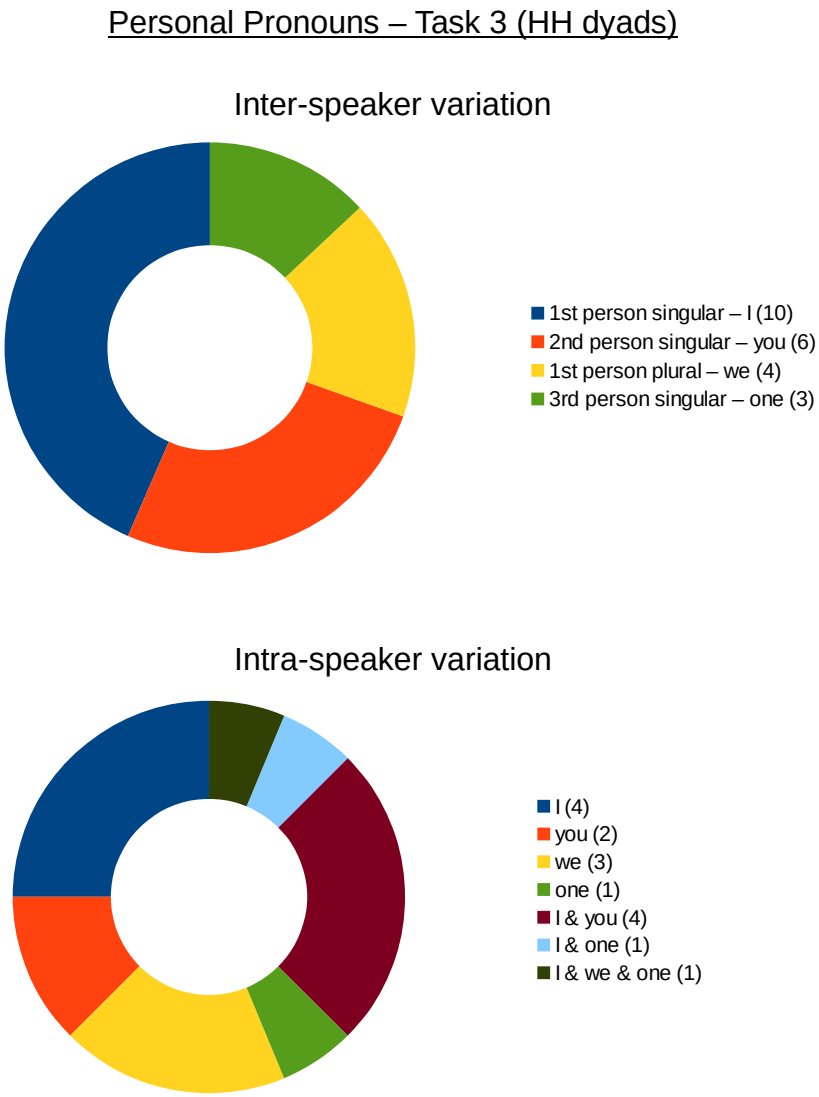

Figure 5.9: Task 3: Personal pronouns uttered between and within tasks (n=16).

Verbal and gestural references to the scene. Six instructors employed deictic gestures to guide the attention of the learner. These pointing gestures were frequently used in combination with indexicals, e.g., "this end" (*dieses Ende*), or verbal references to the scene, e.g., "here" (*hier*). Two instructors raised and exhibited an object to the learner to emphasize its importance. Clark argues that

"placing things just in the right manner" (H. H. Clark, 2003, p.243) is an indicative act in which an object is moved into the addressee's attention. General communicative gestures (e.g., hands poising above objects in the field of attention) were employed by one person and using fingers for counting and raising the index finger when talking about something important were employed by another person.

Verbal references to the visual scene were uttered by all instructors, referring to (i) locations, e.g., "here" (*hier*, *da*), (ii) the manner in which the task was conducted, e.g., "like this" (*so*), or (iii) functioned as disambiguators for information which must be resolved via the visual scene, e.g., "this part" (*dieses Teilstück*), "this green thing" (*dieses grüne Ding*).

Eye gaze. Ten instructors looked at the object or location where the attention was going to be directed before they talked about it. All but one looked at their respective learner before, during and/or after the task. Six instructors frequently looked at the learner during the task description, up to six times.

Task 4 The fourth task is a navigation task. The instructor was giving commands to the learner how to walk to a chair and look whether there was an object. On average, the task duration was 31 seconds (14 sec - 49 sec).

Variation in wording. The edge of the table where the learner was told to start the navigation task was uttered by 14 instructors in four different wordings: "corner" (*Ecke*, *Eck*), "end" (*Ende*), "edge of the table" (*Tischkante*) while three did not explicitly mention the corner of the table. The request to the learner to position her-/himself next to the table was mentioned by 13 instructors uttering 6 different verbs. The action of walking towards the chair was uttered by all instructors in ten different variations, up to five per instructor, e.g., "walk" (*gehen*), "walk around" (*herumgehen*), "pass" (*entlanggehen*, *entlanglaufen*). Also the variation in wording for the action of looking at the chair was high: 17 instructors used 7 different verbs. In contrast, for the condition of the object lying on the chair only one verb was used by 14 instructors, see Table 5.7. One explanation could be that there is a higher variation in wording for processes than for descriptions of situations. However, this hypothesis needs further investigation.

Time Markers. The same instructor who did not use any verbal time markers in Task 2 also did not utter any in Task 4. All other instructors temporarily structured the task verbally (e.g., 'then' (*dann*), 'now' (*jetzt*), 'subsequently' (*anschließend*)).

Table 5.7: Task 4 - Summary of the wording in human-human dyads (n=16). Concepts mentioned by at least 5 participants are listed. ∅ indicates the number of participants omitting a NP for referring to the accordant object.

Object/ Action	**Wording**
Corner	'corner' (*Ecke*) (6), (*Eck*) (1), 'end' (*Ende*) (2), 'edge of the table' (*Tischkante*) (1), ∅ (8)
Table	'table' (*Tisch*) (14), 'tabletop' (*Tischplatte*) (1), ∅ (1)
Chair	'chair' (*Stuhl*) (16), ∅ (6)
Other direction	'other side' (*andere Seite*) (3), (*auf die andere Seite*) (1), 'other direction' (*andere Richtung*) (3), (*anders herum*) (2), 'right side' (*rechts herum*) (2), 'does not matter which one' (*egal in welche*) (1), 'please right' (*bitte rechts*) (1), ∅ (3)
Place	'to place' (*stellen*) (3), (*aufstellen*) (1), (*dahin stellen*) (1), (*dahinter stellen*) (1), (*platzieren*) (1), 'stop' (*stehenbleiben*) (1), ∅ (9)
Walk	'walk' (*gehen*) (13), (*herumgehen*) (11), (*entlanggehen*) (3), (*entlanglaufen*) (1), (*laufen*) (2), (*herumlaufen*) (1), 'walk towards' (*hingehen*) (1), 'pass' (*vorbeigehen*) (1), (*vorbeikommen*) (1), (*vorbeilaufen*) (1)
Look	'look' (*schauen*) (4), (*sehen*) (2), (*gucken*) (2), (*ansehen*) (1), (*blicken*) (1), 'discover' (*entdecken*) (1), 'check' (*nachschauen*) (1), ∅ (5)
Lie	'lie' (*liegen*) (14), ∅ (2)
Stop	'stop' (*stopp*) (5), (*anhalten*) (1), (*halt*) (1), 'no' (*nein*) (2), ∅ (8)
Finished	'finished' (*fertig*) (2), 'that's it' (*das wars*) (2), 'now we are done' (*dann haben wir es geschafft*) (1), ∅ (11)

Compared to the other tasks, only one instructor was taking a step back when finished.

Instructors' perspective. The perspective used by all instructors was 2^{nd} person singular.

Verbal and gestural references to the scene. Most prevalent references to the scene were verbal references in conjunction with deictic gestures. They were used by all instructors. In addition, some used iconic and beat gestures. No verbal references to the scene were uttered without deictic gestures.

Eye gaze. Instructors were frequently looking where the learner needed to go next and they looked at the learner five times on average (two up to seven times). Four finished with a nod towards the experimenter, four with a nod towards the learner, and two first towards the learner and then towards the experimenter.

Human-Robot Dyads

Figure 5.10: Robot learner. An instructor is mounting a tube in a box with holdings.

Questionnaire

None of the six participants has worked with robots before and one had contact with a robot before in a user study. Thus, knowledge about robotics did not influence their interaction. Five out of six participants had the impression that the head of the robot was controlled by an algorithm. Also, one had participated in a study on speech synthesis before and for all others, synthesised speech was new. Overall, the mean value for evaluating the naturalness of the interaction with the robot was 3.33 on a 5 point Likert scale (1 very natural; 5 not natural at all). In general, the majority had the impression that the robot acted autonomously.

Task 3 Task 3 in the human-robot dyads took on average 41 seconds (32 sec - 62 sec), see Figure 5.10 for an example.

Characteristics of spoken language. The characteristics found were similar to those in the HH setting:

- insertions (3) – 'äh', 'the marker the yellow and green one' (*die Markierung das Gelb-gruene*)
- sentence fragments (2) – 'now is done' (*jetzt is fertig*), 'not like this' (*so nicht*)
- repair (1) – 'take assemble' (*nehm stecke*)
- contraction (1)– 'when it' (*wenns*, 'wenn es')
- repetitions (2) – 'assemble it assemble it' (*stecke ihn stecke ihn* – interrupted by verbal feedback of the robot)

- error (1) – 'the' (*das* for 'den')

Variation in wording. Objects relevant for the task are two parts of a tube, two pair of green holdings, two markers, and the right hand of the instructor. Relevant actions are grasping the tube, connecting the parts of the tube, and putting the tube between the green holdings at the two markers. In the human-robot dyads, the following objects and actions were mentioned by all instructors: the tube, the green and yellow marker, the tube mounted in the box, the first green holdings, the second green holdings, the red and yellow marker, taking the tube, assembling the two parts of the tube, and putting it in the box.

However, less variation for all relevant objects and actions occurred than in the human-human interactions, e.g., for "put", see Table 5.8. Part of these results are due to the smaller sample of human-robot dyads. But another part might be due to the adaptation of the human instructors to the robot learner by suppressing intra-speaker variation in wording and employing more prototypical words.

Time Markers. All instructors used verbal time markers to structure the task, e.g., "then" (*dann*), "now" (*nun*). Two instructors took a step back when they had finished the task and three conducted a more restrained posture shift. Only the instructor who did not get any verbal feedback from the robot due to technical problems did not change her posture after finishing the task.

Instructors' perspective. The perspective taken by all instructors in the human-robot dyads was 1^{st} person singular, see Table 5.11. Only one instructor changed from 1^{st} person singular to 2^{nd} person singular after a short break in which she had problems connecting the two parts of the tube. The predominant use of 1^{st} person singular in the human-robot dyads is rather surprising. A possible explanation could be that it is an artefact because the humans did not communicate with the robot preceding the descriptions but they did communicate with the humans and had an opportunity to bond. Another explanation could be that the personal distance between the human and the robot is bigger than between the two humans and the human, therefore, the human does not take the perspective of the robot so easily or that the human doubts that the robot is able to conduct the task and thus does not see the robot in the role of conducting the task. However, these hypotheses were not tested within this work and need further investigation.

Verbal and gestural references to the scene. Five out of six instructors used deictic gestures, and the sixth person moved his fingers which covered the marker

Table 5.8: Task 3 - Summary of the wording in human-robot dyads (n=6). Same concepts as in human-human dyads are listed. ∅ indicates the number of participants omitting a NP for referring to the accordant object.

Object/ Action	**Wording**
Tube	'tube' (*Schlauch*) (4), 'second part of the tube' (*zweite Teil vom Schlauch*) (1), 'loose tube' (*lose Schlauch*) (1), 'a loose one' (*ein Loser*) (1)
Right hand	'right hand' (*rechte Hand*) (4), ∅ (2)
Green and yellow marker	'green and yellow marker' (*grün-gelbe Markierung*) (1), 'yellow and green connection' (*gelb-grüne Verbindung*) (1), 'end with the green and yellow glue' (*Ende mit dem grünen und gelben Kleber*) (1), 'yellow and green marker' (*gelb-grüne Markierung*) (1), 'yellow and green end' (*gelb-grüne Ende*) (3), 'end with the yellow and green marker' (*Ende mit der gelb-grünen Markierung*) (1), 'marker, the yellow and green one' (*Markierung das Gelb-grüne*) (1)
Mounted tube	'tube at the mounting' (*Schlauch bei der Befestigung*) (1), 'mounted tube' (*befestigter Schlauch*) (1), 'tube which is here mounted at the motor' (*Schlauch der hier am Motor befestigt ist*) (1), 'tube' (*Schlauch*) (1), 'pre-assembled tube' (*vorgefertigter Schlauch*) (1), 'other part of the tube' (*andere Teil vom Schlauch*) (1)
First green holdings	'first barrier' (*erste Hindernis*) (1), (*erste Barriere*) (1), 'opening' (*Öffnung*) (1), 'both green separating woods' (*beiden grünen Abtrennhölzer*) (1), 'green marker' (*grüne Markierung*) (1)
Second green holdings	'second barrier' (*zweite Hindernis*) (2), (*zweite Barriere*) (1), 'second opening' (*zweite Öffnung*) (1), 'both green separating walls' (*beiden grünen Trennwände*) (1), 'opposite green marker' (*grüne Markierung gegenüber*) (1)
Yellow and red marker	'red and yellow marker' (*rot-gelbe Markierung*) (1), (*rote und gelbe Markierung*) (1), 'yellow and red marker' (*gelb-rote Markierung*) (1), 'second marker, the red and yellow one' (*zweite Markierung das Rot-gelbe*) (1), 'the other end' (*das andere Ende*) (1), 'red and yellow connection' (*rot-gelbe Verbindung*) (1)
Take	'take' (*nehmen*) (6)
Assemble	'assemble' (*stecken*) (5), (*hineinstecken*) (1)
Put	'put' (*legen*) (4), 'assemble' (*stecken*) (1), 'put through' (*durchlegen*) (1), (*hindurchlegen*) (1), (*durch tun*) (1)
Stay assembled	'is works' (*es geht*) (1), ∅ (5)
Task	'task' (*Aufgabe*) (1), ∅ (5)
Finished	'finished' (*fertig*) (1), ∅ (5)

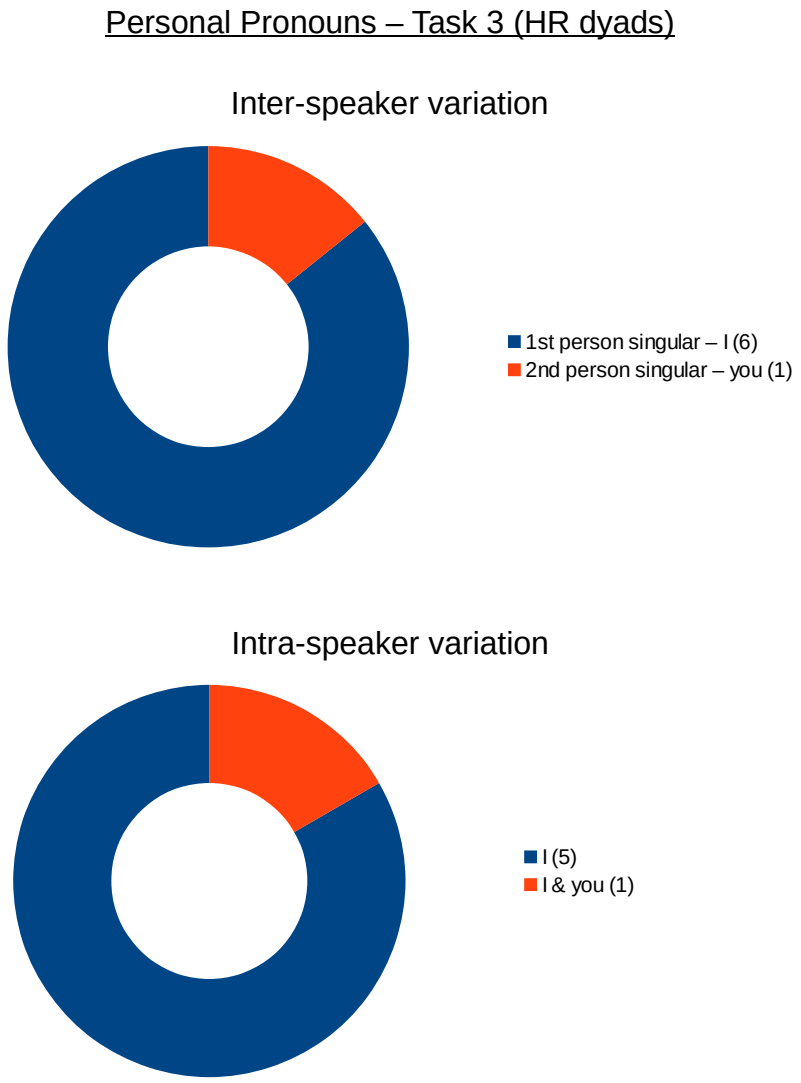

Figure 5.11: Task 3: Personal pronouns uttered between and within tasks (n=6).

before to facilitate the detection of the coloured markers on the tube by the learner.

The gestures were also combined with indexicals, e.g., "this tube" (*diesen Schlauch*), or verbal references to the scene, e.g., "here" (*hier*). Three instructors exhibited the tube or their right hand in order to guide the learner's attention to the object or to their hand.

In addition to gestures, all instructors uttered verbal references to locations, e.g., "here" (*hier*) or disambiguations between objects, e.g., "on this table" (*auf diesem Tisch*).

Eye gaze. All instructors frequently looked at objects before referring to them and all of them looked at the robot learner at the beginning of the task and between 2

Table 5.9: Task 4 - Summary of the wording in human-robot dyads (n=6). Same concepts as in human-human dyads are listed. ∅ indicates the number of participants omitting a NP for referring to the accordant object.

Object/ Action	**Wording**
Corner	'corner' (*Ecke*) (1), (*Eck*) (1), 'edge' (*Kante*) (1), 'frontal edge' (*vordere Tischkante*) (1), ∅ (3)
Table	'table' (*Tisch*) (5), ∅ (1)
Chair	'chair' (*Stuhl*) (5), 'office chair' (*Bürostuhl*) (1)
Other direction	'other direction' (*andere Richtung*) (2), (*anders herum*) (2), ∅ (2)
Place	'to place' (*stellen*) (2), ∅ (4)
Walk	'walk' (*herumgehen*) (3), (*entlanggehen*) (2), (*losgehen*) (1), (*auf etwas zugehen*) (1), (*laufen*) (1), (*herumlaufen*) (1), 'drive' (*fahren*) (2), (*herumfahren*) (2), (*nachfahren*) (1)
Look	'look' (*schauen*) (1), (*sehen*) (1), ∅ (4)
Lie	'lie' (*liegen*) (4), (*drauf liegen*) (4), ∅ (2)
Stop	'stop' (*stopp*) (1), (*stehen bleiben*) (1), (*halt*) (1), 'no' (*nein*) (1), ∅ (2)
Finished	'finished' (*fertig*) (1), ∅ (5)

and 11 times during the task.

Task 4 On average, the task duration was 1 minute and 18 seconds (47 sec - 2 min 12 sec).

Variations in wording. All six instructors mentioned the chair ("chair" (*Stuhl*), "office chair" (*Bürostuhl*)) and the action of walking towards the chair. For the action, 9 different verbs were uttered by the six instructors. The other objects and actions were not mentioned by all instructors, nevertheless up to four different names per object or action occurred, see Table 5.9.

Time Markers. Verbal time markers were uttered during the task description by all instructors ("first" (*als erstes*), "now" (*jetzt*), "then" (*dann*)). After finishing, no instructor took a step back or nodded to signal that the task was done.

Instructors' perspective. The perspective taken by all teachers was 2^{nd} person singular.

Verbal and gestural references to the scene. All instructors used verbal references in combination with deictic gestures. Some additionally employed iconic, pointing or beat gestures without verbal references.

Eye gaze. Instructors again frequently looked at locations before referring to them and they directed their eye gaze towards the learner on average eight times (three to ten times). The robot took longer to navigate to the chair than the human learner, therefore there was more time for the instructor to look at the robot.

5.2.3 Discussion

The information necessary for reproducing the task was transmitted by the instructors via verbal descriptions in combination with exhibiting objects, poising, pointing at objects or locations, placing objects, and eye gazes.

A variety of verbal and non-verbal indicators could be identified, including: (i) at the non-verbal side, communicative acts such as exhibiting, poising, pointing at, placing, gazing and posture shifting, and (ii) at the verbal side of the task descriptions, phenomena such as variations in perspective taking and wording, and of course a broad range of characteristics of spoken language such as repetitions, repairs, occasional shifting from standard variety to dialect.

From the perspective of a robot architecture, the following aspects of human task-based descriptions pose challenges: (i) characteristics of spoken language, (ii) inter- and intra-speaker variations in wording, and (iii) the different perspectives taken by the instructors, and the varying perspectives taken by individual instructors within a task description. Signals which need to be considered to deal with these challenges or to enhance information extraction in general include: (i) verbal references and/or communicative gestures for directing the attention of the learner, (ii) a temporal structuring of the task by verbal means plus a posture shift or a step back when the task was finished, or (iii) eye gaze directing the attention of the learner.

Variation in wording

General results. The analysis of variation in wording shows extensive lexical variation and omitted verbal references for objects and actions. Considering what is visually perceived and what is uttered reveals how differently the same objects and actions are referred to in the utterances. Additionally, the unspoken needs to be grounded in the scene, e.g., the pair of green holdings on the right side of the instructor was not mentioned by five instructors, and the verbal expressions for

spatial perspectives varied, e.g., in Task 3 one instructor named the pair of green holdings on the left side of the instructor "the right one" (*das Rechte*), and another instructor named the same holdings "the left side" (*die linke Seite*).

All this is striking evidence for the importance of vision and for the serious need of multi-modal integration. A robot learning a task from a human instructor has to be able to resolve what is uttered to what is in the world, (see Cantrell et al., 2010).

Differences between tasks. Between, but also within tasks, the commonality of an object or an action may in part influence the variation in wording. While there is no variation for the banana, the strawberry, and the pear in Task 1, objects such as the green barriers in Task 3, which are less prototypical cause more variation in wording. In addition, in case an object needs one or two attributes for disambiguation, this also increases the potential for variation, e.g., the loose and the mounted part of the tube or the pair of green holdings on the right side of the instructor or the pair of green holdings on the left side of the instructor.

Differences between HH and HR interaction. In the human-robot dyads, less variation occurred for all relevant objects and actions than in the human-human interactions, e.g., for "put". This may in part be due to the smaller sample size of the human-robot dyads. It might also be due to the adaptation of the human instructor to the robot learner by trying to utter the information as clearly and as comprehensively as needed. However, for example, for the green barrier at the right side of the instructor, five different nouns, partially with adjectives were uttered. Therefore, the problem of resolving different wordings to a certain entity in the world is also prevalent in human-robot interaction.

There are less omissions of verbal references to objects or actions necessary for conducting the task. For example, for all objects necessary for Task 3 except for the right hand of the instructor, verbal references including a noun are uttered. Here, also an adaptation of the instructor to the robot might occur.

Temporal structuring of the task by verbal means

General results. Except for one participant in Task 2 and Task 4, all instructors uttered time markers in the course of the task description, and instructors frequently took a step back, when finishing the task. The temporal structuring of the task can be used by a robot architecture as a general cue to detect when the task starts and ends, and how the steps in between are temporarily ordered.

Differences between tasks. There were no differences in the use of verbal time markers between tasks. However, in the use of posture shifts, i.e., taking a step backwards when the task was finished, differences occurred. In Task 1, where the instructors had to explain the task towards the camera, four out of 22 participant conducted a posture shift. In the collaborative task, only one stepped back, while in Task 3 where the instructor was explaining and conducting the task, all but two took a step back. In task four where the learner was instructed in how to move to a certain location, again only one instructor conducted a posture shift.

These results show that the degree the instructor and the learner are involved might influence these temporal markers. The more the performance of the task is on the learner-side, the lower is the need to show the learner via a posture-shift that the task is finished. Task 1 is a special case, as there is no learner present.

Differences between HH and HR interaction. The same pattern as for the human-human dyads is visible with regards to temporal structuring.

Variation in the perspective taken by the instructor

General results. Depending on the task, the probability is high that personal pronouns can not be interpreted literally. For example, in Task 3, up to three different personal pronouns were uttered, all referring to the same person, the instructor.

Differences between tasks. In tasks where the learner was involved in conducting the task (Task 2 and Task 4), all personal pronouns referring to agents could be interpreted literally, without exceptions. In Task 1, four different personal pronouns were uttered by all participants ("I", "we", "one", "you"), up to two within task descriptions. However, only "I" was prevalent – it was uttered by half of the instructors. The same personal pronouns were uttered by participants in Task 3. Here, "I" was not prevalent, but uttered only by 1/4 of the instructors and up to three different personal pronouns were uttered, although it was always the instructor who conducted the task. Thus, in tasks where the learner was actively involved, personal pronouns could be interpreted literally, in tasks where only the instructor was conducting the task, the personal pronouns were not reliable and had to be interpreted via the vision system, to identify who is actually conducting the task.

Differences between HH and HR interaction. In the human-robot setting, there was a strong tendency of describing the task in 1^{st} person singular. Even in Task 3, where the robot was only observing, the instructors used "I". Now the

question arises where these differences come from. Before explaining the task to the robot, the participants interacted with human learners in a different task. This might have increased the distance to the robot as a learner and negatively influenced establishing joint attention. Instructors also might have thought of the robot more as a camera than as a learner and thus chose "I" as the preferred perspective.

Verbal references and deictic gestures

General results. Indexicals were uttered by all instructors, referring to (i) locations, e.g., "here", "there" (*hier*, *da*), (ii) the manner in which the task was conducted, e.g., "like this" (*so*), or (iii) objects, where visual information is needed to resolve the ambiguities, e.g., "this tube" (*diesen Schlauch*), "this green thing" (*dieses grüne Ding*). Also, gestures were frequently used in all tasks. However, verbal references in combination with information transmitted via gestures did not suffice to resolve all references.

Differences between tasks. In Task 3, fewer gestures were used than in the other tasks, because instructors often had objects in both their hands and thus could not gesture. Hence, the amount of conducted gestures is also dependent on the task.

Differences between HH and HR interaction. While the use of indexicals was comparable, an increase of gestures occurred in the human-robot settings.

Eye gaze

General results. Task 1 is an exception with regards to eye gaze. In this task, instructors looked at the cheat sheet during the task description and the task was directed towards the camera. In all other tasks, the instructors looked at the referred objects and the majority of instructors looked at the learners in between.

Differences between tasks. In tasks where the learner actively participated, the eye gaze of the majority of instructors was directed at the learner, as opposed to e.g., Task 3, and also the amount of occurrences when the instructor looked at the learner was much higher.

Differences between HH and HR interaction. The instructors looked more often and longer at the learner in the human-robot dyads. This might be explained by the fact that the instructors were not acquainted with robots and in Task 4, it

took the robot longer to navigate, thus the participants had more time to observe the learner.

Overall, it is a considerable challenge to equip robots with system components necessary to understand multi-modal natural human communication. In a task description context, system components and the robot architecture must (i) allow for robust incremental processing of natural speech and of multi-modal communicative signals, (ii) include visual perception of the objects in the scene and the ongoing activity, and (iii) integrate all this in multi-modal representations and the robot's episodic memory.

5.3 Reference Resolution to Objects via Language, Eye Gaze, and Gesture

Results from the previous section have shown, that there is an enormous variation in wording and that verbal referring expressions on their own often do not allow to resolve what is uttered to an action or object in the world. In order to investigate this major challenge of reference resolution in more detail, this chapter focuses on referring expressions to objects. Task 2 and Task 3 were used as the empirical basis for this analysis, as they are the tasks with objects involved as well as an interlocutor. In order to shed light on the variation of human referring expressions in situated task descriptions in the detailed analysis presented in this section, the focus is only on human-human dyads. The reason for this is the assumption that in human-human interaction, humans instructors expect at least the interaction processing capabilities from their human interlocutors as they do from robots. Thus, the variation and spectrum of human referring behaviour is at least as broad as it is in human-robot behaviour. On the other hand, influencing variables in the human-robot dyads are difficult to control and thus it might not be able to generalise the results for other human-robot interaction.

The results from the previous chapter have shown the importance of eye gaze and gesture during situated task descriptions. Also, literature on human situated interaction reviewed in Section 2.1 mention with regards to non-verbal cues mainly eye gaze and gesture. The analysis of this chapter thus investigates referring expressions in situated task descriptions including verbal referring expressions to objects relevant for the task and their interplay with eye gaze and gestures. The general motivation for this section is to identify mechanisms needed to enable a robot to resolve multi-modal referring expressions to objects occurring in natural, situated task-oriented communication and whether it is sufficient to add eye gaze and ges-

tures for the resolution of object references.

5.3.1 Resolution of referring expressions

A referring expression is any noun phrase or representative of a noun phrase referring to an object, an event, an agent, etc. In this analysis, noun phrases, pronouns and spatial indexicals referring to objects are investigated. Referring expression can be multi-modal, as they are often accompanied by non-verbal cues, such as gestures or eye gaze. In order to emphasise which part of the referring expression is meant, "verbal part of the referring expression" is used in the following, when it refers only to the linguistic part. Whether there is a referring expression can be identified via the part of speech. It can be resolved either via one cue (e.g., the participant utters "the green and yellow marker" or utters "the thing" and points at the green and yellow marker) or several cues (e.g., the participant utters "the green and yellow marker" and points at the green and yellow marker). In some cases, multi-modal channels which can serve as a cue in the context might be misleading in another context (e.g., the participant utters "the green and yellow marker" and points at the red and yellow marker). The object intended by the instructor was annotated by two independent annotators based on their competence as interlocutors. "Specific" and "underspecified" in this context refers to whether the references is specific enough to resolve the references via the verbal part of the referring expression.

5.3.2 Research questions

Based on results from the previous section, the following aspects are analysed: (i) variation in the choice of nouns denoting one specific object, (ii) underspecified referring expressions, and (iii) the role of eye gaze and gestures when uttering referring expressions. In the following, the research questions are grouped according to these three aspects.

Variation of expressions referring to one individual object.

RQ1 How high is the inter-speaker variation when referring to an individual object?

RQ2 How high is the intra-speaker variation within one task when referring to individual objects?

Underspecified verbal referring expressions.

RQ3 How often is reference by means of a definite or indefinite noun phrase underspecified and contains neither a description nor a synonym?

RQ4 How often are linguistic referring expressions to objects omitted in the utterance?

RQ5 In natural task descriptions, do initial references to objects always contain a description?

RQ6 In German, how reliable is the gender of a pronoun when looking for an antecedent in the utterance?

RQ7 How often are linguistic antecedents of pronouns omitted?

RQ8 In situated task descriptions, how many pronouns do not refer to objects in the environment?

Multi-modality of referring expressions.

RQ9 How often is eye gaze directed at the intended object when the referring expression is verbally underspecified?

RQ10 How often is a gesture directed at the intended object when the referring expression is verbally underspecified?

RQ11 How often were eye gaze, gestures, and linguistic referring expression misleading, i.e., existent as cues, but directed somewhere else?

RQ12 How many referring expressions could not be resolved via language, eye gaze, and gestures?

5.3.3 Results

In the following, reference resolutions to objects are presented and discussed along the lines of verbal and non-verbal aspects of referring expressions.

Variation of referring expressions per object. In Task 2, the main object to be collaboratively manipulated is the board. For the analysis presented in this section, "the handle" as well as "the coloured marker" refer to the same part oft the board. Thus, the handle is not considered as a separate object in order to avoid a distinction between references where there might be none. The other item referred to in the task is the green and yellow marker.

In Task 3, more objects relevant for the task are involved which also need to be identified by the learner: a loose part of the tube, a mounted part of the tube, the two parts connected to one tube, a green and yellow marker, a yellow and red marker, green holdings at the right side of the instructor, and green holdings at the left side of the instructor.

In natural human-human interaction, variation can also occur due to underspecification or synonyms. In computational linguistics, synonyms are not a problem

as long as they can be handled via lexical databases such as WordNet[7] for English or Universal WordNet[8] for more than 200 languages. For demonstration purposes, the databases were checked whether the different nouns uttered for one object were listed in Universal WordNet. Importantly, no synonyms were found in the database because the different nouns were no synonyms but underspecified NPs.

The verbal part of referring expressions contained specific nouns (e.g., "the board"), underspecified nouns (e.g., "the whole"), pronouns (e.g., "it"), and spatial indexicals (e.g., "here") indicating where to place the object – see Table 5.12 and Table 5.13 for an overview.

Not surprisingly, how instructors refer to objects does not only depend on the object, but also on the task and the other types of objects available for reference. In Task 2, the board is the main object and salient during the whole description. This is also reflected in the number of pronouns occurring as referring expressions. One of the 22 instructors even used pronouns only, without any lexical antecedent when referring to the board, due to its visual salience. The marker is on the handle and the handle as well as the coloured marker refer to the same part of the board. Taken together, referring expressions for the handle and the coloured marker were uttered by 16 (out of 22) instructors.

In Task 3, the seven objects can only be disambiguated via their attributes, not only the noun. For instance, simply saying "the marker" is not sufficient, because the marker can either be green and yellow or red and yellow. Although 59.09% of all noun phrases used as referring expressions contained a noun (e.g., "tube", "marker", and "holdings"), only 10.23% of the referring expressions contained an attribute allowing for disambiguation.

All object references by all instructors taken together for the two objects in Task 2 comprise up to nine different nouns.

In Task 3, up to ten different nouns were used to refer to an individual object, e.g., up to ten for the first pair of green holdings. Considering what is visually perceived and what is uttered reveals how differently the same objects are referred to. The noun phrases varied from very specific (e.g., "green and yellow marker" – *grün-gelbe Markierung* in German) to underspecified noun phrases (e.g., "the thing" – *das Ding* in German). See Table 5.12 for all noun phrases referring to the involved objects in Task 2 and Table 5.13 for all noun phrases referring to the involved objects in Task 3. Only one uttered noun for each object was specific in both tasks (e.g., *Brett*, German for "board"). The other nouns were either no synonyms (e.g., "beam"

[7] http://wordnetweb.princeton.edu/perl/webwn

[8] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/uwn/

Table 5.10: Task 2. For each object, the number of different nouns instructors used to refer to the same object between and within tasks are presented, as well as the number of instructors verbally referring to the accordant object.

	Object	**Noun variation**		**Instructors**
		between tasks	within tasks up to	referring to obj.
Task 2	Board	9	2	22/22
	Green and yellow marker	8	2	18/22

– *Balken* in German) or underspecified (e.g., "the whole' – *das Ganze* in German).

Variations occurred between instructors in the choice of nouns designating one specific object, but also within task descriptions. Up to two different nouns were uttered by one instructor within one task to refer to one object, see Table 5.11.

Table 5.11: Task 3. For each object, the number of different nouns instructors used to refer to the same object between and within tasks are presented, as well as the number of instructors referring to the accordant object.

	Object	**Noun variation**		**Instructors**
		between tasks	within tasks up to	referring to obj.
Task 3	Tube	4	2	14/16
	Loose part of the tube	7	2	15/16
	Mounted part of the tube	5	1	15/16
	Green and yellow marker	6	2	15/16
	Yellow and red marker	2	2	12/16
	First green holdings	10	2	16/16
	Second green holdings	8	2	16/16

Table 5.12: Task 2. Summary of the object references uttered by all 22 instructors. The first number in brackets indicates the number the referring expression is uttered all together, the second number indicates the number of instructors uttering the referring expression.

Object	**Referring expressions - noun phrases**
Board	'board' (*das Brett*) (12;9), 'object' (*das Objekt*) (3;2), (*dieses Brett*) (2;2), 'thing' (*das Ding*) (2;2), 'the whole' (*das Ganze*) (2;2), 'item' (*diesen Gegenstand*) (2;1), (*das ganze Brett*) (1), (*dieses Ding*) (1), (*das Teil*) (1), (*dieses Teil*) (1), 'beam' (*diesen Balken*) (1), (*Balken*) (1), 'arrangement' (*die Anordnung*) (1), 'device' (*das Gerät*) (1)
Green and yellow marker	'yellow and green marker' (*die gelb-grüne Markierung*) (1), 'marker, the yellow and green one' (*die Markierung, an die Gelb-grüne*) (1), 'marker' (*die Markierung*) (1), 'green and yellow markers' (*die grün-gelben Markierungen*) (1), 'marker of the yellow and green one' (*die Markierung von dem Gelb-grünen*) (1), 'handle' (*der Griff*) (3;3), (*dieser Griff*) (2;2), (*der Henkel*) (2;2), (*der gelb-grüner Henkel*) (1), (*diese Hantel*) (1), 'thing' (*dieses Teil*) (1), 'green and yellow boarder' (*die grün-gelbe Umrandung*) (1), 'lever with the yellow and the green colour' (*der Hebel mit der gelben und der grünen Farbe*) (1), 'the side' (*die Seite*) (1), 'here' (*hier*) (1), ∅ (4)

Table 5.13: Summary of the object references uttered in Task 3 by all 16 instructors. The first number in brackets indicates the number the referring expression is uttered all together, the second number indicates the number of instructors uttering the referring expression.

Object	**Referring expressions - noun phrases**
Tube	'tube' (*der Schlauch*) (11;9), (*ein Schlauch*) (1), (*Schlauch*) (1), 'the whole' (*das Ganze*) (2;2), 'the tube thing' (*dieses Schlauchteil*) (1), 'the appendant parts' (*die zugehörige Teile*) (1), 'the two tubes' (*die zwei Schläuche*) (1), 'the connected tube' (*der verbundene Schlauch*) (1), ∅ (5)

Table 5.13: (continued)

Object	Referring expressions - noun phrases
Loose part of the tube	'tube' (*der Schlauch*) (5;5), (*dieser Schlauch*) (2;2), 'loose pipe' (*das lose Rohr*) (1), 'the part of the tube' (*das Teil von dem Schlauch*) (1), 'the end-piece' (*das Endstück*) (1), 'one end' (*das eine Ende*) (1), 'the one tube' (*der eine Schlauch*) (1), 'the part/thing' (*das Teil*) (1), 'this other tube' (*dieser andere Schlauch*) (1), 'this side' (*diese Seite*) (1), ∅ (4)
Mounted part of the tube	'tube' (*der Schlauch*) (2;2), 'tube' (*dieser Schlauch*) (1), 'pipe' (*das Rohr*) (1), (*dieses Rohr*) (1), 'segment of the tube' (*dieses Teilstück des Schlauches*) (1), 'end' (*das Ende*) (1), 'second tube' (*der zweite Schlauch*) (1), ∅ (8)
Green and yellow marker	'green and yellow marker' (*die grün-gelbe Markierung*) (5;4), (*diese grün-gelbe Markierung*) (1), (*die gelb-grüne Markierung*) (5;4), (*diese gelb-grüne Markierung*) (1), 'green and yellow end' (*das grün-gelbe Ende*) (3;3), 'marker' (*die Markierung*) (2;2), 'end where the green and yellow is attached' (*dieses Ende wo das Grüne und das Gelbe dran ist*) (1), 'end with the yellow and green marker' (*das eine Ende mit der gelb-grünen Markierung*) (1), 'yellow and green connection' (*die gelb-grüne Verbindung*) (1), 'green and yellow part' (*der grün-gelbe Teil*) (1), 'green and yellow section' (*der grün-gelbe Abschnitt*) (1), 'this side' (*diese Seite*) (1), 'green thing' (*das grüne Teil*) (1), ∅ (1)
Yellow and red marker	'red and yellow marker' (*die rot-gelbe Markierung*) (5;5), (*die gelb-rote Markierung*) (1), 'the yellow and red one' (*der Gelb-rote*) (2;2), (*die Rot-gelbe*) (1), 'marker' (*die Markierung*) (1), 'red marker' (*die rote Markierung*) (1), 'where it is yellow and red' (*wo es gelb-rot ist*) (1), 'the red one' (*das Rote*) (1), ∅ (4)

Table 5.13: (continued)

Object	Referring expressions - noun phrases
First green holdings	'mounting' (*diese Befestigung*) (1), 'this side' (*diese Seite*) (1), 'holding' (*die Halterung*) (1), '(right) first holding' (*der erste Halter*) (1), (*unsere erste Halterung*) (1), (*die rechte erste Halterung*) (1), 'first barrier' (*das erste Hindernis*) (1), 'green thing' (*dieses grüne Ding*) (1), 'two blocks' (*diese beiden Klötze*) (1), 'right green marker' (*diese rechte grüne Markierung*) (1), 'right channel' (*der rechte Kanal*) (1), 'appliance' (*diese Vorrichtung*) (1), ∅ (5)
Second green holdings	'second holdings' (*die zweite Halterung*) (3;3), (*der zweite Halter*) (1), 'other green holdings' (*die andere grüne Halterung*) (1), 'holdings' (*die Halterung*) (1), 'other channel' (*der andere Kanal*) (1), 'other appliance' (*die andere Vorrichtung*) (1), 'these two' (*diese Beiden*) (1), 'side' (*die Seite*) (1), 'left side' (*die linke Seite*) (1), 'second green thing' (*dieses zweite grüne Ding*) (1), 'second barrier' (*das zweite Hindernis*) (1), ∅ (4)

Underspecified verbal referring expressions. In the following, the verbal part of all referring expressions, the verbal part of initial references, and pronoun resolution will be discussed.

Verbal part of referring expressions.[9] When taking a closer look at referring expressions for the board in Task 2, only 22.82% can be resolved based on linguistic information alone, see Figure 5.12. 16.30% contain the noun *Brett* and 6.52% are pronouns with proximate, congruent, and specific antecedents.

Regarding the seven objects in Task 3, fewer pronouns were uttered because these objects were less salient during the task than the board in Task 2. However, in Task 3, a larger amount of noun phrases was underspecified due to the need to disambiguate not only via nouns, but also via adjectives. Also the salience of individual objects frequently varied. Altogether, 36 pronouns were uttered to refer to the seven objects in Task 3, but only two of these pronouns had a proximate, congruent, and specific antecedent.

[9] In this current section, a verbal referring expression is resolvable if it contains both the necessary noun (e.g., "marker") and the necessary adjectives (e.g., "green and yellow") for disambiguation. Noun phrases such as "the green and yellow" (*das Grün-gelbe*) are not taken into account in this analysis. However, they will be considered in next section on the interplay of linguistic forms and non-verbal modalities in object references.

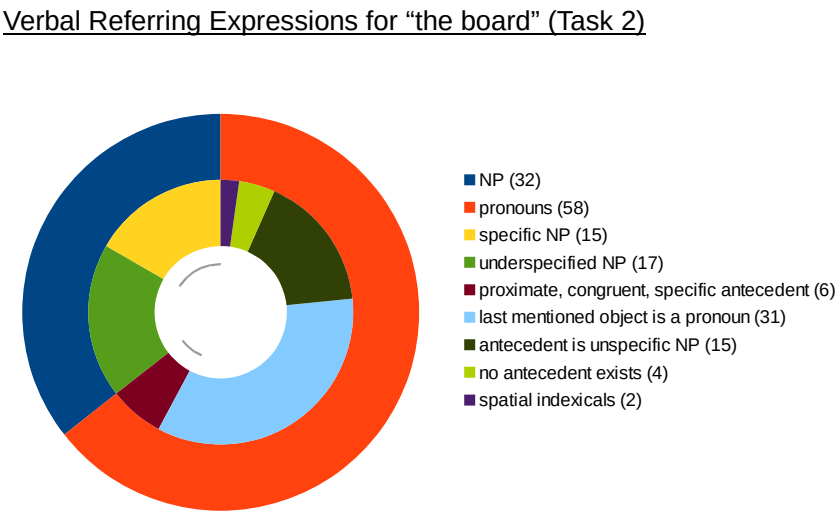

Figure 5.12: All in all, 32 noun phrases, 58 pronouns and two spatial indexicals were uttered as referring expressions for the board in Task 2 by 22 instructors. 22.82% can be resolved only using language, indicated by grey lines.

Spatial indexicals “here” (*hier, da*) for referring to objects were mainly uttered when the object was either mounted to, or part of another object, such as the mounted tube, the holdings, and the markers. *Hier, da* was not uttered for self-contained objects, such as the board and the loose part of the tube.

One case occurred where the linguistic referring expression was misleading: one instructor in Task 3 referred to the green and yellow marker as “the green thing” (*das grüne Teil*). This referring expression would be more appropriate for one of the two green holdings than for the marker. However, the human learner was still able to reproduce the task.

Verbal part of initial references. When objects were mentioned for the first time in a task, subjects used not only noun phrases with a uniquely identifiable object in the visual scene, but also noun phrases underspecified for conceptual content, spatial indexicals, pronouns, and some instructors even omitted a linguistic referring expression for a specific object during the whole task.

The first initial reference in Task 2 for the board contained the noun “board” by eight (out of 22) instructors. Nine instructors used underspecified noun phrases, four uttered a pronoun and one a spatial indexical. Out of these 22 instructors, four omitted a linguistic referring expression for the “board” for a long time. For instance, one instructor who used a pronoun when first talking about the board started the task description with “please grasp with your left hand the handle, **I with my right hand** we lift it [...]” (*fasst du bitte mit deiner linken Hand an den Griff* ***ich mit meiner rechten Hand*** *wir heben es hoch [...]*). All in all, only 36.36% of the initial

references for the board contain sufficient verbal information for disambiguation. Considering all initial references for the board and the marker, only ten out of 40 contain sufficient verbal information for disambiguation, see Figure 5.14). In the majority of the cases, the referents can be resolved by extralinguistic means only.

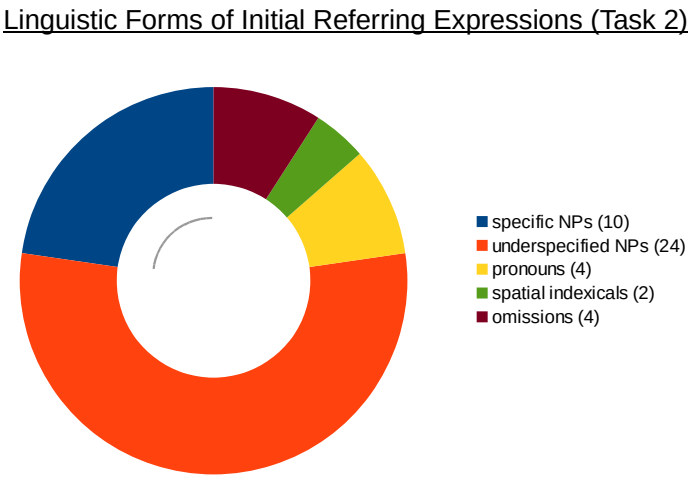

Figure 5.13: Verbal part of initial references. The pie chart contains all first mentions of the two objects in Task 2. Only initial references by means of specific NPs can be resolved on a linguistic basis only, indicated by a grey line. Omissions of initial references refer to objects, for which no noun is uttered at all by an instructor during the whole task.

In Task 3, more objects are involved and some of them can only be differentiated by colour or spatial relation. Only 13.46% of all initial references for the seven objects provide sufficient verbal information for disambiguation, see Figure 5.14. In both tasks, some instructors omitted referring expressions for objects.

Table 5.15: Task 3 - All initial references to objects by all 16 participants.

Object	**Initial references to objects**
Tube	'tube' (*der Schlauch*) (4), (*ein Schlauch*) (1), 'the two tubes' (*die zwei Schläuche*) (1), 'the connected tube' (*der verbundene Schlauch*) (1), 'the whole' (*das Ganze*) (1), 'the appendant parts' (*zugehörige Teile*) (1), es (4), ihn (1), ∅ (2)

Table 5.15: (continued)

Object	**Initial references to objects**
Loose part of the tube	'tube' (*der Schlauch*) (5), (*dieser Schlauch*) (2), 'the one tube' (*der eine Schlauch*) (1), 'this other tube' (*dieser andere Schlauch*) (1), 'the part of the tube' (*das Teil von dem Schlauch*) (1), 'loose pipe' (*das lose Rohr*) (1), 'the end-piece' (*das Endstück*) (1), 'the other one' (*der andere*) (1), 'it' (*den*) (1), (*das*) (1), ∅ (1)
Mounted part of the tube	'tube' (*der Schlauch*) (2), (*dieser Schlauch*) (1), 'second tube'(*der zweite Schlauch*) (1), 'segment of the tube' (*dieses Teilstück des Schlauches*) (1), 'pipe' (*das Rohr*) (1), (*dieses Rohr*) (1), 'end' (*das Ende*) (1), 'the other one' (*das andere*) (1), 'here' (*hier*) (5), (*da*) (1), ∅ (1)
Green and yellow marker	'green and yellow marker' (*die grün-gelbe Markierung*) (2), (*die gelb-grüne Markierung*) (3), (*diese grün-gelbe Markierung*) (1), (*diese gelb-grüne Markierung*) (1), 'yellow and green connection' (*die gelb-gruene Verbindung*) (1), 'green and yellow end' (*das grün-gelbe Ende*) (3), 'green thing' (*das grüne Teil*) (1), 'green and yellow part' (*der grün-gelbe Teil*) (1), 'end where the green and yellow is attached' (*dieses Ende wo das Grüne und das Gelbe dran ist*) (1), 'side' (*die Seite*) (1), ∅ (1)
Yellow and red marker	'where it is yellow and red' (*wo es Gelb-rot ist*) (1), 'marker' (*die Markierung*) (1), 'red and yellow marker' (*die rot-gelbe Markierung*) (5), (*die gelb-rote Markierung*) (1), 'the yellow and red (section/part)' (*der gelb-rote (Abschnitt/Teil)*) (2), 'red marker' (*die rote Markierung*) (1), 'this end' (*dieses Ende*) (1), ∅ (4)
First green holdings	'holding' (*die Halterung*) (1), 'green thing' (*dieses grüne Ding*) (1), 'appliance' (*diese Vorrichtung*) (1), '(right) first holding' (*der erste Halter*) (1), (*die rechte erste Halterung*) (1), 'this side' (*diese Seite*) (1), 'first barrier' (*das erste Hindernis*) (1), 'this marker' (*diese Markierung*) (1), 'right channel' (*der rechte Kanal*) (1), 'right green marker' (*diese rechte grüne Markierung*) (1), 'here' (*hier*) (4), (*da*) (2)

Table 5.15: (continued)

Object	Initial references to objects
Second green holdings	'second holdings' (*die zweite Halterung*) (2), 'this mounting' (*diese Befestigung*) (1), 'second green thing' (*dieses zweite grüne Ding*) (1), 'other appliance' (*die andere Vorrichtung*) (1), 'holdings' (*die Halterung*) (1), (*der zweite Halter*) (1), 'left side' (*die linke Seite*) (1), 'second barrier' (*das zweite Hindernis*) (1), 'other channel' (*der andere Kanal*) (1), 'side' (*die Seite*) (1), 'other green holdings' (*die andere grüne Halterung*) (1), 'here' (*hier*) (3), (*da*) (1)

Pronoun resolution

The pronouns used were either the personal pronoun "it" (*es*) or the demonstrative pronouns "it" (*das*) and "this" (*diese*). In order to resolve a pronoun with current computational models, the pronoun has to be congruent with the antecedent (i.e., match in number and gender) and occur in certain proximity.

However, in Task 2, three instructors used pronouns for referring to the board where the gender of the pronoun was not congruent with the gender of the antecedent (3, 1, and 8 pronouns did not match the gender). In these cases, either the pronoun *das* or *es* were applied as "default" pronouns referring to something unspecific (e.g., "the thing" – *das Ding* in German). In German, the reliability of the gender of a pronoun when looking for an antecedent depends on the gender of the antecedent. In case the gender did not match, the pronouns *das* and *es* were employed referring to an unspecific neuter antecedent. If the gender of the antecedent is neuter, the gender of the pronoun always matched (e.g., for "board" *das Brett*). If it was feminine or masculine (e.g., for "tube" *der Schlauch*) the gender of the pronoun did not always match.

For example, 44.83% of all pronouns referring to one of the three parts of the tube did not match the gender of their antecedent.

Multi-modality of referring expressions. With regard to **gestures**, pointing gestures were employed by 16 (out of 22) instructors at least once in Task 2, see Figure 5.15, and 6 (out of 16) in Task 3 and frequently used in combination with indexicals (e.g., "this end" – *dieses Ende* in German), especially spatial indexicals (e.g., "here" – *hier* in German), see Figure 5.16.

Referring expressions were accompanied by pointing gestures in 13.27% of all referring expressions in Task 2 and 10.23% in Task 3. Gestures were the only cue

Table 5.14: Task 2 - All initial references to objects by all 22 participants.

Object	**Initial references to objects**
Board	'board' (*das Brett*) (4), (*dieses Brett*) (2), 'beam' (*diesen Balken*) (1), 'this thing' (*dieses Ding*) (1), (*dieses Teil*) (2), 'item' (*diesen Gegenstand*) (1), 'object' (*das Objekt*) (1), 'the whole' (*das Ganze*) (1), 'this' (*das*) (4), ∅ (5)
Green and yellow marker	'yellow and green marker' (*die gelb-gruene Markierung*) (1), 'marker' (*die Markierung*) (2), 'green and yellow markers' (*die gruen-gelben Markierungen*) (1), 'marker of the yellow and green one' (*die Markierung von dem gelb-gruenen*) (1), ∅ (17)
Handle	'handle' (*der Griff*) (4), (*dieser Griff*) (2), (*der Henkel*) (1), (*der gelb-grüner Henkel*) (1), (*diese Hantel*) (1), 'green and yellow boarder' (*die gruen-gelbe Umrandung*) (1), 'lever with the yellow and the green colour' (*der Hebel mit der gelben und der grünen Farbe*) (1), 'here' (*da*) (1), ∅ (10)
Table	'table' (*der Tisch*) (20), (*Tisch*) (1), ∅ (1)

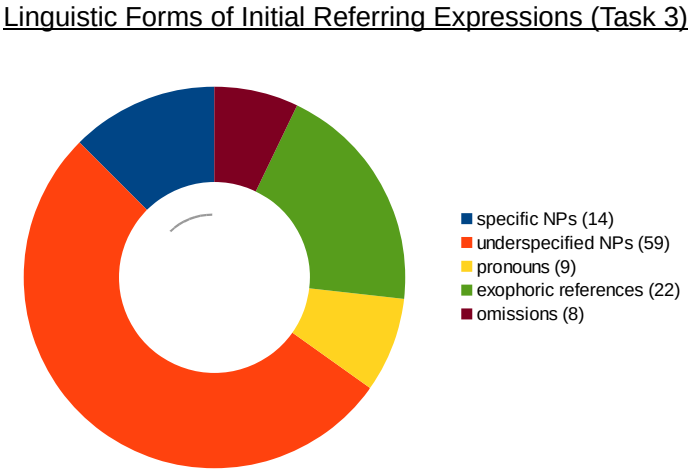

Figure 5.14: Verbal part of initial references. The pie chart contains all first mentions of the seven objects in Task 3. Only initial references by means of specific NPs can be resolved on a linguistic basis only, indicated by a grey line. Omissions of initial references refer to objects, for which no noun is uttered at all by an instructor during the whole task.

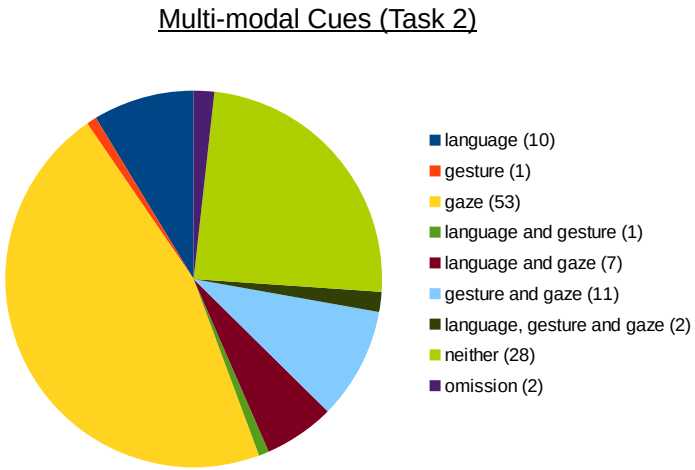

Figure 5.15: Multi-modal cues of all referring expressions in Task 2.

to resolve a reference in 0.88% in Task 2 and 2.84% in Task 3.

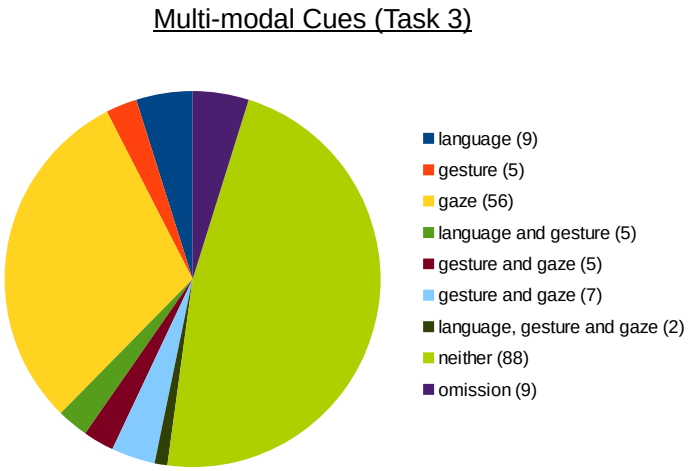

Figure 5.16: Multi-modal cues of all referring expressions in Task 3.

However, referring expressions where the instructor pointed somewhere else were also uttered. In case the instructor used deictic hand gestures, they were directed towards the referred object in 65.22% of the cases in Task 2. In the other 34.78% of the cases, the location of where to place the salient object was fixated. In Task 3, only one instructor once pointed somewhere other than the verbally referred object.

In the psycholinguistic literature, it is emphasized that **the gaze of the speaker** is an important cue for disambiguation (e.g., see Prasov & Chai, 2008; Hanna &

Brennan, 2007; Knoeferle & Crocker, 2006). In Task 2, in 64.60% of all referring expressions, the eye gaze was directed at the referred object. In 46.90%, the verbal part of the referring expression was underspecified, no pointing gesture was used and only the eye gaze of the instructor was directed at the object. In Task 3, gaze was directed at the referred object in 39.77% of all cases and was the only cue in 31.82% of all referring expressions.

Language was rarely misleading, but often underspecified. An exception is one instructor who denoted the green and yellow marker with "the green thing" (*das grüne Ding*). This description would be more appropriate for the green holdings. In both tasks, two cases occurred where the reference could only be resolved via pronoun resolution.

These results show that with regard to **the multi-modal interplay**, eye gaze was often the only modality referring at the intended object. Different from gestures, the instructor's eye gaze is not applied only for certain references, but is present the whole time. Hence, depending on the context, learners might track the path of the instructor's eye gaze as it is moving along and follow it to the location where a salient object needs to be placed.

Using language (specific noun phrases), pronoun resolution, deictic gestures, and eye gaze as cues for the resolution of references, it is possible to resolve 75.22% of all referring expressions (excluding omissions) to the two objects in Task 2 and 50% to the seven objects in Task 3.

5.3.4 Discussion

Results discussed in the previous section revealed a vast variation in wording when participants referred to objects and action. The analysis of the role of language, eye gaze, and gestures, in this section for referring expressions to objects in Task 2 and Task 3 has shown the following:

Variation of expressions referring to one individual object

The high variation in wording showed the deviation from what is perceived and what is uttered. Noun phrases ranged from very specific (e.g., "the green and yellow marker") to very general, lexically underspecified concepts (e.g., "the thing"). Depending on the number of objects involved as well as the similarity between objects, the requirements on verbal referring expressions differ. In case of mainly one object (as in Task 2), its visual salience increases. In case of more objects, the verbal referring expression needs one or two attributes to distinguish them, it is more difficult

to refer to them solely via language. This is also reflected in the data: including anaphora resolution, 21.24% of all verbal referring expressions can be resolved via language in Task 2. In Task 3, only 11.36% can be resolved solely via language.

Verbal referring expressions either contained specific nouns (e.g., "the board" or "the green and yellow marker"), nouns lexically underspecified for conceptual content (e.g., "the thing"), pronouns (e.g., "it", "this"), or spatial indexicals (e.g., "here", "there").

Inter-speaker variation when referring to an individual object. Instructors used up to ten different nouns to refer to one specific object in Task 3, and up to nine in Task 2.

Intra-speaker variation within one task when referring to individual objects. Within tasks, up to two different nouns were uttered to refer to the same object. No lexical entrainment occurred, as the learner was mainly observing.

Underspecified verbal referring expressions

Amount of underspecified noun phrases. For the board in Task 2, only 16.30% contain the noun *Brett* and 6.52% are pronouns with proximate, congruent and specific antecedents. In Task 3, a larger amount of noun phrases was underspecified, because disambiguation of all relevant objects at least one additional adjective is needed, e.g., "the green and yellow marker" versus "the red and yellow marker".

Omitted verbal referring expressions. In both tasks, some instructors omitted verbal referring expressions for objects. For the board in Task 2, all participants uttered a verbal reference, however, four omitted it for a long time, where the board already played a major role.

The linguistic form of initial references to objects. In Task 2, only 25% of initial references to relevant objects contained sufficient verbal information for disambiguation. For relevant objects in Task 3, it was even worse with 13.46%.

Pronoun resolution and congruent gender in German. In German, the reliability to resolve a pronoun by its gender depends on the gender of the antecedent. In case the gender did not match, the pronouns were neuter and the antecedent was not. In Task 2, three participants uttered pronouns where the gender of the

pronoun did not match the gender of the antecedent. In Task 3, 44.83% of all pronouns referring to one of the three parts of the tube did not match the gender of their antecedent. In this context, grammar-based pronoun resolution is deemed to fail and extra-linguistic information is required for reference resolution.

Omitted antecedents in pronoun resolution. When uttering a verbal referring expression for the board in Task 2 for the first time, four instructors used a pronoun. In Task 3, nine pronouns were uttered by all 16 instructors to refer to relevant objects for the first time.

Multi-modality of referring expressions

Eye gaze and the resolution of underspecified noun phrases. Eye gaze was directed at the referred object in 64.60% of all referring expressions in Task 2. In 46.90%, it was the only cue referring at the intended object. In Task 3, gaze was directed at the referred object in 39.77% of all cases and in 31.82%, it was the only cue.

Pointing gestures and the resolution of underspecified noun phrases. The frequency of verbal referring expressions for which gestures were the only cue referring at the intended object was much lower than for eye gaze: 0.88% in Task 2 and 2.84% in Task 3.

Contradiction of language, eye gaze, and gesture. In case the reference was resolvable via language, it was never misleading. Pointing gestures were directed at the referred object in the majority of cases. In some cases, they were directed at locations where the referred object had to be put. However, although eye gaze is a very prevalent cue in the data, it has to be treated with caution. In Task 2, 24.88% are not resolvable via language, eye gaze and gesture and 50% in Task 3. In these cases there is eye gaze – but not at the intended object. The difficult question is how to distinguish, when eye gaze can be used as a cue for a robot architecture and when it can not be used.

Amount of referring expressions not resolvable via language, eye gaze and gesture. Specific noun phrases, pronoun resolution, pointing gestures, and eye gaze refer to the intended object in 75.22% of all referring expressions to the (two) objects in Task 2 and 50% to the (seven) objects in Task 3.

Now the question arises, which additional cues are needed to resolve the remaining referring expressions, and also how to deal with eye gaze as a potential cue for object references.

Additional non-verbal cues might include:

- *Salience* – check the visual field and the discourse for a currently salient object
- *The object currently manipulated* – check whether the instructor or both the instructor and the learner hold an object in their hands
- *Preceding eye gaze* – in case the eye gaze of the instructor is directed at the learner, the last object the instructor looked at is extracted
- *Flexible pronoun resolution* – check whether the last mentioned referring expression is a pronoun which has a resolvable antecedent
- *The attribute* – in case the adjective of the referring expression matches the attribute of one object, it can be used as a cue; in case it matches more objects, it can still constrain target objects
- *The noun* – in case the referring expression is underspecified, the employed noun can still constrain the target object (e.g., in Task 2 there are two markers)
- *Proximity* – in case a person is asked to manipulate an object and the referring expression refers to two objects (e.g., two handles), the probability is higher that the referring expression refers to the object reachable for the person (see also Kruijff et al. (2010); Hanna & Tanenhaus (2004))

In the next section, additional non-verbal as well as verbal cues are extracted in order to resolve all occurring references. The third analysis focuses on Task 3, as it involves more objects than Task 2. The interplay of linguistic forms and non-verbal modalities is further investigated, shedding more light on the reliability of different non-verbal cues.

5.4 The Interplay of Linguistic Forms and Non-verbal Modalities in Object References

The previous section has shown the small percentage of referring expressions which can be resolved via language. Gaze and gesture are frequently directed at the inferred object and can thus be used for the resolution of references. However, the reliability of eye gaze as a potential cue needs to be further investigated, as it is always directed somewhere, but it can not always be used for reference resolution. Language, eye gaze, and gesture are not sufficient to resolve all referring expressions to objects. In about 1/4 of all referring expressions in Task 2 and 1/2 of all referring

expressions in Task 3, additional cues are needed. In this Section, all references to objects in Task 3 are investigated. For each reference, the linguistic form in combination with non-verbal information necessary for resolving the reference is extracted.

In this section, not only references to relevant objects, but to all objects physically present in the shared environment are investigated in order to be as comprehensively as possible. In the 16 human-human dyads, 205 object references were uttered. There are no deictic gestures without verbal references to objects, thus, language is the primary cue whether there is a reference to an object or not.

A closer look at the data showed the potential of including the object(s), the instructor grasped last and is currently holding as well as the object(s), the instructor grasped before the last one(s) and is still holding. Thus, the annotation was extended by a tier including this cue, see Section 5.1.5.

5.4.1 Research questions

This Section investigates the following research questions:

RQ1 Which additional cues are needed to resolve all object references in situated task descriptions?

RQ2 Is there a correlation between linguistic form and non-verbal cues?

RQ3 How reliable is eye gaze for reference resolution in a task description where different objects are involved?

5.4.2 Results

All in all, 205 object references were uttered to the loose part of the tube, the mounted part of the tube, the two parts of the tube connected, the green and yellow marker, the red and yellow marker, the pair of green holdings on the right side of the instructor, the pair of green holdings on the left side of the instructor, the motor block, and the round table. Language only covers 46 references, i.e., 22.44% via uniquely identifiable noun phrases (42) and pronoun resolution (4) via a congruent, proximate, and uniquely identifiable antecedent. Uniquely identifiable noun phrases include examples such as "the red and yellow marker" (*die rot-gelbe Markierung*) but also "the red and yellow section" (*der rot-gelbe Abschnitt*) and even "the red one" (*das Rote*) if there is no other object with the attribute "red".

In general, 22.44% can be resolved via the discourse. Now the interesting question is how can ALL of those 205 references be resolved? This is where additional non-verbal cues come into play.

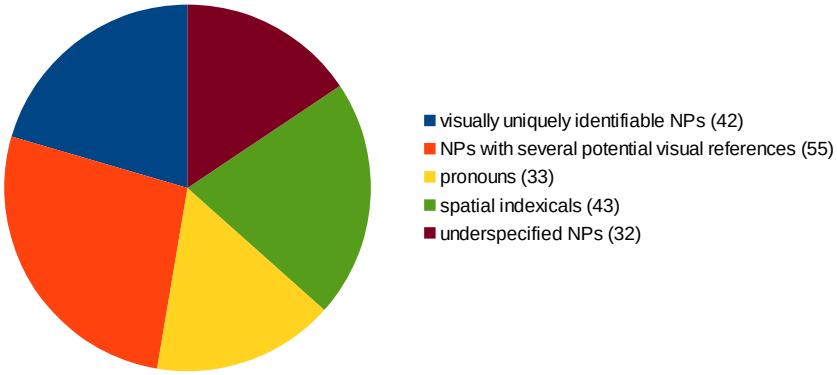

Figure 5.17: This figure shows all linguistic types of references in Task 3. The types include visually uniquely identifiable NPs, NPs with several potential visual references, e.g., "marker" or "tube" which refer to more than one object in the scene, pronouns, spatial indexicals, such as "hier" or "da", or underspecified NPs. Underspecified NPs refer to examples such as "the thing" (*das Ding*), but also to NPs with a similar semantic concepts such as "the channel" (*der Kanal*) for a pair of green holdings.

Depending on the type of verbal reference, see Figure 5.17, different non-verbal cues are relevant. With regard to gestures, the following types of gestures were used: (i) emblems, e.g., thumbs-up, (ii) beats (see McNeill, 1992), (iii) pointing, (iv) poising (over the referred object or location), and (v) exhibiting gestures (where the instructor holds up an object for the interaction partner to see). With regard to object references, the last three types of gestures are of interest. Clark (1996) argues that gestures are considered composite parts of references made with deictic expressions.

In Task 3, the instructors often had both hands occupied when they connected or mounted objects and thus had no hand free to point. In this context, poising and exhibiting gestures are of equal importance as pointing gestures to direct the attention of the interlocutor at a certain object or location. In this respect, all three gestures are deictic acts.

Half of the participants made use of these three deictic gesture types: all in all they employed 16 pointing, six exhibiting and three poising gestures. For the majority of pointing gestures, eye gaze was preceding the pointing gesture, followed by a verbal referring expression. In 12 of the 16 cases where pointing was employed, and in two out of the six exhibiting gestures, the verbal reference did not allow to uniquely identify the object referred to. While the gesture indicated to the listener where to direct attention, the instructor did not verbally utter the nouns of referred objects. Eye gaze was not misleading in these cases with deictic gestures, although

in some cases the instructor looked at the listener in between. However, eye gaze is less reliable than gestures, as eye gaze (in contrast to gestures) is always present and verbal referring expressions are predominantly insufficient to resolve verbal referring expressions.

However, language, gaze, and gesture are not sufficient for reference resolution in a situated task description and further cues will be investigated in the following.

Resolution of noun phrases

The majority of uttered references to object are NPs: 62.93%. In the following, they are clustered in (i) visually uniquely identifiable NPs, (ii) NPs with more than one potential visual reference, and (iii) underspecified NPs, such as general concepts (e.g., "the thing") and similar semantic concepts (e.g., "the channel").

Visually uniquely identifiable noun phrases. Verbal references of visually uniquely identifiable objects (20.49%) is the easiest case for object reference resolution. In case a NP is uttered including the noun as well as the attributes (if there are any necessary for disambiguation) of one visually observable object, the references can be resolved. There was no occurrence of an utterance including a visually uniquely identifiable object and the instructor intended to refer to another object (again, based on the evaluation of the annotators).

Noun phrases with several potential visual references. NPs with several potential visual references refer to more than one potential object in the scene and thus additional visual cues are needed for reference resolution, e.g., "the tube" (*der Schlauch*) or "the green part" (*das grüne Teil*). Out of these 55 references, 28 can be resolved, if **the object the instructor grasped last** is added as a cue. 12 out of these 28 utterances are additionally accompanied by gaze and two by gaze and gesture. For the resolution of references, pointing, exhibiting, and poising gestures are taken into account, as they all direct the attention of the learner at a certain object.

In addition to these 28 references, 5 can be resolved via **gaze and gesture** of the instructor directed at the referred object.

Gesture and holding an object are very reliable cues in this context (i.e., seldom misleading). Information about the object, the instructor grasped last and is currently holding is only misleading in three cases, which means that in three cases a noun (e.g., "tube") was uttered an the tube the instructor was holding did not refer to the tube he or she uttered. These three cases are special and occurred

in the following contexts: they were uttered (i) during a summary before starting the task "Now it is about mounting a tube" (*Hier geht es darum, dass wir einen Schlauch verlegen*) (ii) when the tube was already mounted, thus the knowledge is needed, that after connecting two objects of the same type (e.g., two tube), the tube referred to after assemblance is the connected tube, and (iii) at the beginning, when two objects were manipulated: "you take this tube and this other tube" (*du nimmst diesen Schlauch und diesen anderen Schlauch*) - the instructor was holding the tube referred to first, thus the second tube can not be the one the instructor is holding.

As opposed to the information which object the instructor grasped last and is currently holding, the gesture of the instructor is not only relevant for NPs with several potential visual referents, but also for all other NPs, pronouns and spatial indexicals. For resolving NPs with several potential visual references, it serves as an important cue seven times. For all object references, gesture can be used to resolve the object reference 28 time. However, in addition to these 28 occurrences, in nine cases it cannot be used to resolve references, as it is misleading and directed somewhere else. For these nine occurrences, gesture as a relevant cue can be avoided by looking at the following aspect: (i) is the temporal sequence of the gesture so long, that it lasts on during several object references, and (ii) is a person gesturing with two hands in two different directions.

Thus, before using the information if the instructor is holding an object or using a deictic gesture as a cue for reference resolution, these aspect need to be investigated:

(i) Whether the utterance is a **summary of the task before starting the task description in detail**: this aspect can be investigated by looking at the words at the beginning of the utterance and the beginning of the task description, e.g., sentences starting with "now it is about" (*hier geht es darum*) or "the task consists of three steps" (*die Aufgabe besteht aus drei Schritten*). These utterances also probably contain verbs such as "is about" (*geht darum*) or "consits of" (*besteht aus*).

(ii) **Semantic knowledge about verbs**: for example, when two objects of the same type are assembled or connected, the separate objects are not referred to anymore (e.g., if there are two tubes, before their assemblance, they need to be disambiguated, but after their assemblance, there is only one tube although its parts are still visible). Another example is that for certain verbs a hand of the person conducting the action is involved, e.g., for "take" (*nehmen*) in "you take this tube and this other tube" (*du nimmst diesen Schlauch und diesen anderen Schlauch*) the first and the second uttered "tube" refer to different objects and

for both a hand is needed. In case a cue such as gesture or information about the object the instructor grasped last can not refer to only one object twice, additional cues need to be included, e.g., which object the instructor grasps next. In three cases, a deictic gesture starts too early and in three cases the gesture still lasts on during other object references. Thus, the gesture needs to be tracked and if a gesture lasts on during more than one object reference, semantic knowledge about the verb can be used to resolve the reference.

(iii) **Identify parallel gestures**: One instructor gestured with both hands at two objects in parallel, but this only occurred once and spanned three object references. In this case, the moving hand was the one directed at the relevant object and the other one was for keeping the attention also at another object.

However, after adding gesture and information about the object, the instructor grasped last as an additional cue, there are still 22 references unresolved. Another important cue is **semantic knowledge about the verb in combination with visual information about where a certain object moves or which other object or location it touches**. Referring to NPs with several potential visual references, there is already a pre-selection of visually perceivable objects. If you have, for example, the utterance "insert with the right hand in this pair of holdings" (*mit der rechten Hand in die Halterung einführen*), you have the knowledge that during an "inserting"-action two objects touch. The right hand of the instructor is moving towards or touching a certain object and that is the object needed for object resolution. This cue is relevant for 15 object references out of these 55. The missing seven reference can be dealt with by (i) identifying whether the utterance is a **summary of the task before starting the task description in detail** - these object references can be ignored as it might not always be possible to link the mentioned object to already existing objects in the scene, (ii) the knowledge that **after assembling two objects of the same type, they are not referred to separately anymore**, e.g., if "tube" is uttered after assembling the two tube, it refers to the assembled one and not to the separate parts anymore, (iii) the knowledge that **two colours mentioned one after the other refer to one object**, if there is an object with this attribute, e.g., "the green and the yellow one" (*das Gruene und das Gelbe*), and (iv) the knowledge that **a person can do something with the right and the left hand in combination with knowledge about the verb**, e.g., that the utterance "and then you put with the left one" (*und dann tust du mit der Linken*) refers to the left hand.

Gaze is very often not directed at the referred object and therefore no reliable cue on its own.

Underspecified NPs. In Task 3, 32 underspecified NPs are uttered by the 16 instructors. This group of object references can be divided in NPs lexically underspecified for their conceptual content, such as "the thing" (*das Ding*) and NPs with a similar conceptual content.

19 verbal references were uttered containing a general concept, such as "the whole" (*das Ganze*), "this end" ((*dieses Ende*), "the other part" (*das Andere*) or "this side" (*diese Seite*).

13 verbal references were uttered containing a noun which does not fully match the referred object but a similar semantic concept. Examples are "pipe" (*Rohr*) or "channel" (*Kanal*) instead of "tube" (*Schlauch*) or "mount" (*Halter*) instead of "holding" (*Halterung*).

Important information to allow reference resolution of these underspecified NPs is (i) to identify an utterance as a task summary four times, (ii) the gestures of the instructor seven times, (iii) knowledge about the verb in combination with information about the object, the instructor grasped last 13 times, as well as (iv) knowledge about the verb in combination with visual information about where a certain object moves or which other object or location it touches eight times.

Resolution of pronouns

Out of these 33 pronouns, five can be resolved via discourse, via a proximate, congruent and specific antecedent, e.g., in "I take the green and yellow end of the tube and connect it [...]" (*ich nehme das gruen-gelbe Ende des Schlauches und verbinde es [...]*), see Table 5.16 for an overview.

As pronoun resolution via discourse fails in the majority of cases, additional cues are needed. As opposed to NPs, for pronouns the object the instructor grasped before he/she grasped the last object and is still holding is a more important cue than the object he/she grasped last. In the data, the reference could be resolved **via the object the instructor was "still holding"** eleven times. In six cases, the reference referred to **the object, the instructor grasped last and there was no object he/she was "still holding"**. Nine times the instructors were assembling and holding the two parts of the tube while they already referred to it as a whole. Thus, **knowledge about the verb** is needed in order to know that after combining two objects, it can be referred to as one. Only in one case, holding and still holding were misleading: "you would have to insert the tube with the right hand in the pipe, insert no that is somehow, and when it is then inserted [...]" (*du müsstest den Schlauch mit der rechten Hand reinstecken am Rohr reinstecken nein das ist irgendwie und wenn er dann drinnen steckt [...]*). This is a rare case of a

Table 5.16: Summary of all pronouns and spatial indexicals referring to visually present objects by all 16 instructors in Task 3 are listed.

	Object	**Referring expressions**	
		pronouns	spatial indexicals
Task 3	Tube	20	0
	Loose part of the tube	10	1
	Mounted part of the tube	0	13
	Green and yellow marker	3	2
	Yellow and red marker	0	1
	First green holdings	0	14
	Second green holdings	0	6
	Round table	0	2
	Motorblock	0	4

summary of an already described process and needs to be identified via the verb. The object which is inserted first is the same object as the one which is in the next step already inserted. It also includes a sentence fragment of a meta-description when the instructor commented that it is not working the way he wanted it. In another case an instructor was holding no object and it was during a summary before starting the task description in detail. Thus, for pronoun resolution, the following cues are needed:

- to identify a summary before the actual task description
- distinguish between utterances that describe the task and meta-descriptions referring to the performance (e.g., "I am not very good in doing this" (*das kann ich nicht so gut*), "that is the task" (*das ist die Aufgabe*), "it is a bit difficult" (*es geht ein bisschen schwer*)
- before using visual cues or pronoun resolution, knowledge about verbs is

needed: if two parts of a similar type are assembled, the pronoun or noun uttered after that refers to the assembled object. If there are, for example, two very similar objects such as the loose part of the tube and the fixed part of the tube and a verb for assembling the two objects, such as "connect" (*stecken, anstecken, hineinstecken, zusammenstecken* or "combine" (*verbinden, kombinieren*), it takes two objects. It is also possible to utter "one connects that", then one already refers to the assembled object. In the other cases, where the pronoun takes two objects or an object and a location, they are separate during the assembling action, but immediately after the assembling action, there is a new object: the assembled object. Additionally, knowledge is needed that e.g., objects move during "put" (*legen*) or "take" (*nehmen*) to/from a certain location.

- information about the object the instructor grasped before the last object he/she grasped
- information about the last object he/she grasped

11 additional pronouns could be resolved via pronoun resolution, but for these pronouns, the accordant antecedent had to be resolved via visual cues.

Resolution of spatial indexicals

For spatial indexicals, holding is often misleading and thus not very reliable as a cue. The most important cue for spatial indexicals is **the combination of knowledge about the verb, whether there is a pause before of after "here" (*hier, da*) and towards which object the already mentioned argument of the verb moves / which it touches**. This occurred 41 times, two times accompanied by deictic gestures. In the majority of the cases, the already mentioned argument of the verb is still moving towards the object, but in some task descriptions, the verbal description is a bit slower than conducting the action. In these cases, the last movement of two objects towards each other can be used as a cue. One instructor omitted the verb "and now this end through here" (*und jetzt dieses Ende noch hier durch*), still "here" can be identified as a location, "this end" can be resolved via visual cues and moves towards the left pair of green holdings, thus, the reference can be resolved. There are three spatial indexicals uttered by the instructors during **summaries before actually conducting the task**.

To resolve spatial indexicals in general, the following information is necessary:

- is the utterance a summary before the actual task description.

- knowledge about the verbs is a necessary prerequisite: "you have to put this tube here" (*du musst den Schlauch hier reinstecken*), "you have to take this tube here" (*du musst den Schlauch hier nehmen*). Due to the knowledge how many arguments a verb takes or allows (the argument structure of a verb), it can be determined that the first "hier" refers to another object / location, the second to the same object as the preceding noun. Another example refers to objects which are parts of other objects, such as the red and yellow marker, which is part of the tube. The verb "insert" (*einführen*) for example takes two objects. If the instructor talks about the red and yellow marker and is holding the tube, there needs to be an additional object, where the instructor puts the tube, because the marker and the tube count as one object referring to the verb.

- in case there is a NP immediately preceding or following the spatial indexical, it is important whether there is a pause before or after "here".

- based on knowledge about the verb, it is important where the already resolved arguments of the verb move or if they moved immediately before and are now touching an object.

It can also be observed that spatial indexicals refer more often to fixed objects (e.g., the markers on the tube, the pair of green holdings or the fixed part of the tube) and very seldom to loose objects (e.g., the loose part of the tube, or the connected part of the tube). It is only used once to refer to the loose part of the tube, and in that case, the instructor refers successively at the two parts of the tube and it is not definitely clear which tube is meant by which referring expression. It is also not essential to understand the description: "Ok first we take this tube and this one here" (*also nehmen wir zuerst den Schlauch und den hier*).

Holding is not very reliable for reference resolution of spatial indexicals, as locations are often not touched.

5.4.3 Discussion

The analysis revealed, that in addition to verbal referring expressions, eye gaze and gestures, the following cues are needed to resolve all verbal references to objects in Task 3:

- *The object the instructor grasped last and is currently holding.* This information was in particular important to resolve noun phrases with several potential

visual referents. Out of 55 verbal referring expressions with two or more potential referents, 28 can be resolved if this visual cue is added. However, this information is less important to resolve references of pronouns and spatial indexicals.

- *The object the instructor grasped before the last one and is currently holding.* For the resolution of pronouns, this cue is valuable for reference resolution. Only in cases where there was no object the participant grasped before the last one, the pronoun could be resolved via the object he/she grasped last.

- *Knowledge about the verb.* Knowledge about the verb is important for the resolution of noun phrases, pronouns, and spatial indexicals. In some cases, a gesture was directed at one object while two are uttered (e.g., "You have to take this tube here" (*Du musst den Schlauch hier nehmen*) versus "You have to insert this tube here." (*Du musst den Schlauch hier einfügen*)). In order to distinguish, if there is one object mentioned twice or if there are references uttered for two distinct objects, information about the number of arguments a lexical item takes need to be used.

- *A pause before or after spatial indexicals.* In case a NP or a pronoun was uttered preceding or following a spatial indexical, a pause could be used to distinguish, whether it referred to the same object as the pronoun or NP, or to another object or location.

- *Visual information about where a certain object moves or which object or location it touches.* This was important information to resolve NPs and spatial indexicals. When two objects are brought together to execute a certain action, not all objects are grasped. For example, if a marker is put in a pair of holdings, the holdings are not grasped but the marker moves towards and touches the holdings.

- *Identify parallel gestures.* One participant gestured with two hands in two different directions. He was hoovering above a certain area to keep the attention of the learner also there, while he grasped a tube. In this case, the gesture moving towards an object was the one that could be used to resolve the reference.

- *Identify a summary of the task at the beginning of the task description.* This cue is needed in order not to use visual cues for reference resolution.

- *Identify meta-descriptions referring to the performance.* This information is also needed in order not to resolve references to objects in these utterances (e.g., "it is a bit difficult" (*es geht ein bisschen schwer*).
- *General knowledge*:
  - *After assembling two objects of the same type, they are not referred to separately.* When two parts of a tube are connected, the connected object is then the new "tube".
  - *Two colours mentioned one after the other refer to the same object, if there is an object containing the mentioned colours.* If the reference "the green one and the yellow one" (*das Grüne und das Gelbe*) are uttered, it refers to one, not to two objects, in case there is an object containing both attributes.

The analysis also revealed a **correlation between linguistic form and non-verbal cues**. Information, which object the instructor grasped last is, for example, necessary to resolve object references containing noun phrases. On the other hand, information about the object the instructor grasped before the last one and is still holding is an important cue for pronoun resolution. These results show the tight linkage between the discourse and the interaction.

With regard to **the reliability of eye gaze** the data show that eye gaze is not an adequate cue to resolve references to objects in Task 3. The reason for this might also be due to the setup of the task. Participants frequently looked at the object they intended to refer to before they uttered a verbal referring expression. However, they also frequently looked at objects they currently manipulated together with the object they referred to. They checked whether the manipulation action they just conducted was still the way they intended it to be (e.g., if the tubes are still connected and did not fall apart), or at the location where they had to put that object in the next step. Thus, although the eye gaze often predicted the upcoming area of attention, it is not a reliable cue for reference resolution in tasks, where there are different objects and locations involved and they are manipulated one after the other, without a break.

5.5 Summary and Discussion

In the last three sections, the empirical data was investigated with three different foci. Nevertheless, each analysis builds upon the results of the preceding analyses, including the first one, which builds upon the results of the pilot study.

In Section 5.2, an **explorative analysis of inter- and intra-speaker variation** was presented. The aim of this analysis was to identify (i) general patterns of human multi-modal task descriptions that could be used by a robot to extract task-relevant information, as well as (ii) potential challenges for robots in situated multi-modal task descriptions. Based on the results of the pilot study, the following aspects of human situated multi-modal task descriptions were extracted:

Characteristics of spoken language. In all tasks, characteristics of spoken language occur, which need to be identified and dealt with.

Variation in wording. There was extensive variation in wording between and also within tasks. The analysis reveals extensive lexical variation and omitted verbal reference for objects and actions. Pronouns frequently lacked an antecedent in language and needed to be resolved via the visual context, and also intra-speaker variation played a role in wording.

Instructors' perspective. In tasks in which the instructor was conducting the task alone while describing it, participants varied a lot in the personal pronouns they used, between, and again also within tasks. Uttered personal pronouns varied not only between but also within speakers and task descriptions. In tasks where the learner was actively involved and had to follow instructions, instructors uttered only personal pronouns for agents which could be literally interpreted ("I", "you", "we").

Time markers. Verbal time markers were frequently uttered and can be used by the learner to structure the task.

Verbal and gestural references to the scene. Gestures and indexicals referring to locations, the manner in which a task was conducted, or objects were frequently produced.

Eye gaze. In all tasks except for Task 1 (there was no learner present and participants frequently looked at a cheat sheet) instructors often looked at the referred objects and the leaner.

The main differences in how the task was transmitted in HH and HR settings are an increase of gestures in the human-robot setting and a strong tendency of describing the task in 1^{st} person singular. Instructors did not interact with the robot before the task (as opposed to the human learner) which might have increased the distance to the robot as a learner. However, the number of eye gazes toward the robot learner was on average higher than toward the human learner.

These results underpin the importance of non-verbal communication cues in human task descriptions and the need to incrementally incorporate linguistic and visual information in a robot architecture to resolve referents of unspecific noun phrases or pronouns lacking verbal antecedents. The major challenge for robots in this context is the variation in wording. Although language is very important in structuring the task and to reliably identify that the instructor is referring to an object, for the resolution of referring expressions itself, the role of language is minor.

Due to these result, references to objects were further investigated in Section 5.3. It was analysed, how often information transmitted via language, eye gaze and gestures refers to the object intended by the instructor in Task 2 and Task 3. The results of the analysis showed the following aspects:

The variation of expressions referring to one individual object. Verbal referring expressions contained either specific nouns, underspecified nouns, pronouns, or spatial indexicals. Variation when referring to an individual object occurred between and within tasks.

Underspecified verbal referring expressions. In both tasks, a large amount of verbal referring expressions was underspecified. Additionally, verbal referring expressions were omitted and could only be referred to visually. Only a small amount of initial referring expressions could be resolved via language. Also pronouns could rarely be resolved via language, as antecedents were often omitted and the gender of the pronouns was often not congruent with the gender of the antecedent.

The multi-modality of referring expressions. Eye gaze was directed at the referred object a bit more than half of the cases in Task 2 and in a bit less than half of the cases in Task 3. However, it has to be treated with caution, because in the other half of the cases, in could not be used as a cue to resolve the references and was misleading. Pointing at objects was a reliable cue in the majority of cases, but not very frequently used, as for example in Task 3 the hands of the instructor were occupied most of the time. In case a verbal referring expression was resolvable via language, it was never misleading.

Although many references could be resolved via language, eye gaze and gestures, about 1/4 of the referring expressions in Task 2 and about 1/2 of the referring expressions in Task 3 could not be resolved via language, eye gaze, and gestures.

In order to extract missing cues needed to resolve these remaining references, the analysis in Section 5.4 was conducted. In this section, all verbal referring expressions

are systematically investigated according to their linguistic form and the additional cues needed to resolve all of them.

The verbal referring expressions are grouped in (i) visually uniquely identifiable noun phrases, (ii) noun phrases with several potential visual referents, (iii) underspecified noun phrases (NPs lexically underspecified for their conceptual content, or NPs with a similar conceptual content), (iv) pronouns, and (v) spatial indexicals.

In order to resolve all of these expressions, the following cues are needed:

- Language
- The objects in the visual field
- Gesture
- The object the instructor grasped last and is currently holding
- The object the instructor grasped before the last one and is currently holding
- Knowledge about the verb
- A pause before or after spatial indexicals
- Visual information about where a certain object moves or which object or location it touches
- Identify parallel gestures
- Identify a summary of the task at the beginning of the task description
- Identify meta-descriptions referring to the performance
- General knowledge

  After assembling two objects of the same type, they are not referred to separately.

  Two colours mentioned one after the other refer to the same object, if there is an object containing the mentioned colours.

Eye gaze turned out to be a not very reliable cue in a task description, where different objects and actions are relevant. The results also show a correlation between linguistic form and the additional cues. The object, the instructor grasped last and is currently holding is a very important cue to resolve NPs with several potential referents. For pronouns on the other hand, the object the instructor grasped before the last one and is still holding is more important. Only if there is no object the instructor is still holding, the object the instructor grasped last can be used for reference resolution. However, an important cue for all linguistic forms is, for example, knowledge about the verb such as argument structure.

In the discussion of the preceding analysis, the following potentially relevant non-verbal cues for reference resolution were mentioned: the salience of an object, the object currently manipulated, preceding eye gaze, flexible pronoun resolution,

the attribute or the noun in the verbal part of the referring expression, or the proximity of objects and agents. While the visual salience, the flexible pronoun resolution, and resolving objects only via its adjective or noun play an important role in the third analysis, preceding eye gaze as well as the proximity of objects play a secondary role. However, these cues should also be kept in mind as depending on the communication context and task, they could also play a primary role.

In order to develop reference resolution mechanisms for robots, not only the above mention cues are relevant – some cues are only relevant for some linguistic forms – but also the order in which information from the verbal and non-verbal cues is extracted.

In general, the findings of the first explorative analysis also show differences between tasks: (i) the variation in wording decreases if objects are selected with rather unambiguous names, e.g., banana, (ii) the instructors' perspective is unambiguous if not only the instructor, but also the learner is involved in the performance, while it can not be literally interpreted and varies also within tasks, if only the instructor performs, (iii) deictic gestures to the scene decrease, if the instructor needs both hands for conducting the task, but exhibiting and poising gestures might incease, (iv) eye gaze is more often directed at the learner, if it is a collaborative performance, and (v) time markers and disfluencies occur in all tasks.

The results of the second and the third analysis are general findings for reference resolution to objects in situated task descriptions and will also transfer to different tasks.

In the following Chapter, the results of all three analyses are used in order to formulate challenges for robot architectures and resulting design suggestions for human-robot interaction.

Chapter 6

Challenges and Architectural Design Suggestions for Robots in Multi-modal Human-Robot Interaction

The previous chapter has shown a broad spectrum of variation in human multi-modal task descriptions, between and within tasks. In order to enable a robot to deal with this vast variation, concrete challenges and resulting lessons for agent design will be formulated in this section. First, the different challenges will be introduced in Section 6.1 – two with regard to the verbal part of referring expressions and two with regard to multi-modal cues. Subsequently, in Section 6.2 design suggestions will be made in order to develop artificial agents that can deal with the extracted principles of situated, multi-modal task descriptions. The chapter will conclude with a summary and discussion in Section 6.3.

6.1 Challenges for robots in situated task descriptions

Our findings are in line with converging evidence that human communication is inherently multi-modal (H. H. Clark & Krych, 2004; Brennan, 2000; H. H. Clark & Brennan, 1991) and thus provide additional information that is critical for robot architecture design. Due to the detailed analyses based on the research questions of the previous chapter, challenges can be extracted and suggestions provided to develop a multi-modal reference-resolution mechanism for situated multi-modal task

descriptions. The first two challenges concern solely language, the third and the fourth concern verbal referring expressions to objects and their interplay with other cues necessary for reference resolution.

6.1.1 The verbal part of referring expressions

The majority of information transmitted via language is underspecified and insufficient for an artificial agent to resolve references to visually perceived physical objects. Rather, the instructors' additional gestures, knowledge about the verb (such as argument structure) or object related actions etc. will have to be taken into account as well. For example, Gundel et al. (2012); Dahan et al. (2002); Gundel et al. (1993); Almor (1999) assume that the salience of potential referents is related to the focus of attention on certain entities in the discourse situation.

Challenge 1 - Variation of noun phrases when referring to one specific object

During an interaction, participants have to permanently adapt to each other in order to interpret utterances of the interlocutor (Barr & Keysar, 2006). Especially with regards to noun phrases, there is a large variation of content words (see Furnas et al., 1987, 1984; Brennan, 1996).

In a study by Brennan (1996) on lexical variability in human-human dialogue, the probability that between trials instructors used the same word for a specific object was only 10%. Within trials, however, variability was relatively low and lexical entrainment occurred.

In case the interaction is situated and cues other than language are at hand to refer to objects, the variation even increases.

In the data presented in this thesis, the instructors largely vary expressions for referring to a single object not only between, but also within tasks. A possible explanation might be that in the instructor-learner dyads one person was talking and explaining the task while the other one was mainly listening and, therefore, no lexical entrainment between speakers could occur.

In the data, three different groups of underspecified noun phrases occurred:

- noun phrases with several potential referents, e.g., "marker" is uttered, if there is a green and yellow marker, and a red and yellow one;
- noun phrases lexically underspecified for conceptual content, e.g., "the thing", or "the whole";

- noun phrases with a similar conceptual content, e.g., "beam" is uttered for "board".

Regarding synonyms, for each object in Task 2 and Task 3, no synonyms were uttered according to WordNet, i.e., all variations besides the specific NPs are underspecified NPs, pronouns and spatial indexicals. Still, this variety of expressions has to be mapped to one entity in the situated environment.

Challenge 2 - Resolution of pronouns

For interpreting pronouns, the negotiation of meaning between interlocutors also plays a major role.

Personal pronouns – variation in perspective. Personal pronouns ("I", "we" or "you" etc.) can be used in many languages such as German and English as impersonal pronouns transmitting structural knowledge (Kitagawa & Lehrer, 1990). This is also reflected in the data. In tasks, where the learner was involved in conducting the task (Task 2 and Task 4), all personal pronouns could be interpreted literally. However, in tasks, where only the instructor was performing and explaining to the learner a sequence of actions to be conducted, a variety of different personal pronouns was used. Participants even changed the personal pronouns they used in the course of the action up to two times.

The amount of spatial indexicals referring to objects was even higher than the number of pronouns and their reference can only be resolved in combination with visual cues.

Pronoun resolution – referring expressions to objects. In this respect, different sub-challenges occurred:

Pronouns lacking verbal antecedents. Although some model on reference resolution account for non-verbal cues such as gestures and eye gaze (e.g., Gundel et al., 2012), they are very vague on how these non-verbal cues should be integrated with language. The results presented above have shown that only a very small amount of pronouns uttered in Task 2 and Task 3 can be resolved solely via the discourse.

Pronoun resolution in German when the gender of the pronoun and the antecedent are not congruent. Studies have shown that gender and accessibility information influence referent consideration during initial processes of pronoun

resolution (see Arnold et al., 2000). However, in German, the gender of the pronoun and the gender of the antecedent often do not match.

6.1.2 The need for cues in addition to language

The results of the above presented analyses have shown, that the number of referring expressions resolvable solely via the discourse is rather small. In Task 3, only 22.44% of 205 referring expressions could be resolved via language.

Studies conducted by Brennan et al. (2008) and Lozano & Tversky (2006) also revealed that speech (as opposed to non-verbal communicative cues) has the potential to inhibit communication. H. H. Clark & Krych (2004) argue that for certain types of communication visual reference resolution is faster and more reliable. The results presented in this thesis also show that underspecified referring expressions in language have to be resolved mainly via visual cues. For example, instructors employ pronouns as initial references to a particular object. In order to resolve the pronoun to an element of the domain of interpretation, information about where different visual cues are directed need to be included for reference resolution.

Challenge 3 - reliability of different cues for reference resolution

Although humans could use only language to transmit information relevant for a certain task, various multi-modal communication cues such as eye gaze and gestures are applied when showing and explaining a task to a learner, especially when the learner is physically co-present. Eye gaze and gestures are the cues dominantely mentioned in literature as additional cues for situated reference resolution (see McNeill, 1992; Kendon, 2004; H. H. Clark & Krych, 2004; Hanna & Brennan, 2007). Some authors also mention other cues such as exhibiting, poising, placing, and orienting objects, head nods and head shakes (e.g., H. H. Clark & Krych, 2004).

Speaker's **eye gaze** may function as an indicator for upcoming utterances (see e.g., Frischen et al., 2007). The results of the presented analyses show that gaze and gestures were frequently directed at the referred object. However, gaze was not very reliable, as it referred not at the intended object but somewhere else for about half of the references. This might be due to the setup of Task 3. Participants looked back and forth between objects they currently manipulated, objects, they already manipulated before, and locations, where objects had to be put in the future. Thus eye gaze is not a very reliable cue for reference resolution mechanisms for robots.

With regard to *gestures*, not only pointing gestures, but also poising and exhibiting gestures are deictic acts and transmit information relevant for the resolution of

references. In Task 3, gesture is a reliable cue in the majority of cases and rarely misleading.

Other cues that could be extracted based on the empirical data are (i) the object the instructor grasped last and is currently holding, (ii) the object the instructor grasped before the last one and is still holding, (iii) knowledge about the verb (such as argument structure), (iv) a pause before or after spatial indexicals (v) visual information about where a certain object moves or which object or location it touches, (vi) the identification of parallel gestures, (vii) the identification of a summary of the task at the beginning of the task description, (viii) the identification of a meta-descriptions referring to the performance, (ix) knowledge that after assembling two objects of the same type, they are not referred to separately anymore, and (x) knowledge that two colours mentioned one after the other refer to the same object, if there is an object containing the mentioned colours. While the first eight cues are rather general for situated task descriptions, the last two are specific for the objects and actions occurring in Task 3.

Challenge 4 - interlinkage of verbal referring expressions and additional cues for reference resolution

Some theoretical and computational approaches to reference resolution take also non-verbal cues such as the eye gaze of the interlocutor, the visibility of objects in the shared environment, and/ or pointing gestures into account (e.g., Gundel et al., 2012; Kehler, 2000; Chai et al., 2006; Huang & Mutlu, 2014). However, the interlinkage between language and non-verbal modalities is still either unclear or too vague for situated task descriptions. Also, current models for reference resolution do not include additional verbal and non-verbal cues for references resolution, except pointing gestures and eye gaze.

The analyses presented in this thesis have shown that the linguistic form determines the importance of non-verbal cues for references resolution and according to each linguistic cue, the sequence in which information is extracted differs.

6.2 Lessons for Agent Design

A main finding from the situated task interactions is that as language takes more of a secondary, scaffolding role. It becomes less informative than information transmitted via visual channels.

Variation of noun phrases when referring to one specific object

In case of underspecified noun phrases, either (i) NPs lexically underspecified for conceptual content were uttered, such as "the thing", "the object", (ii) nouns carrying similar semantic information as the denoted object without being synonyms (e.g., "beam" for "board"), or (iii) NPs with several potential referents, such as "tube" if there is a loose tube and a tube mounted to another object.

In case of *NPs lexically underspecified for conceptual content*, solely extra-verbal cues are needed for reference resolution. The only role of language is the information that there is – at this point in time – a reference to an object.

For *nouns (or verbs) carrying similar semantic information* as the object referred to (e.g., "beam" for "board"; "barrier" or "obstacle" for "holdings") a design suggestion would be to explicitly include meaning negotiation in the agent design. For example, a verbal referring expression containing the noun "beam" is uttered. However, the system is not able to detect a beam in its visual field. Independent of the method used for semantic representations (e.g., ontologies, representations based on vectoral semantics etc.) the system can then check whether objects in the visual field are very similar objects to the one, mentioned by the interlocutor. In case it detects, for example, that "board" is very close to "beam", the reference can still be resolved. The data also shows an (even broader) variation of uttered verbs for one action. This strategy can also be applied for reference resolution of verbs.

NPs with several potential referents can be automatically identified. To disambiguate between them, e.g., between two different tubes, additional cues need to be consulted. The object the instructor grasped last and is currently holding is especially relevant in this context.

Underspecified noun phrases also raise the issues of learning new representations. New objects and new wordings for already known objects can only be learned via experience, i.e., extensive exposure to a combination of linguistic and visual input. Only via extensive exposure, the usage of NPs lexically underspecified for conceptual content (e.g., "the thing") can be learned and *wrong* wordings such as NPs with a similar conceptual content can be ignored (e.g., "pipe" for "tube").

The resolution of pronouns and spatial indexicals

In addition to noun phrases, pronoun resolution and the resolution of spatial indexicals are a great challenge for human-robot interaction. The majority of pronouns cannot reliably be resolved by means of linguistic information and this might hinder the correct interpretation of utterances, especially when disconnected from visual

information.

Personal pronouns referring to agents can often not be interpreted literally. Based on the results of this data, the following design suggestion was extracted: as soon as the learner knows, who is conducting the task, personal pronouns with regard to who is conducting the task can be interpreted: In case the learner is actively involved in conducting the task, all personal pronouns referring to agents can be interpreted literally. In case the instructor is conducting the task on his/her own while explaining it, personal pronouns referring to agents can be substituted by "I" or "one".

In situated task descriptions, antecedents of pronouns are often omitted or verbally underspecified. Although, e.g., the Givenness Hierarchy by Gundel (1993) includes eye gaze and gesture, it does not explicitly say how these cues should be interlinked. However, the analyses presented in this thesis also show, that it is not only eye gaze and gestures, but other cues which should be interlinked with language.

Also, in German, gender is not a reliable cue for linguistic reference resolution. Neuter is frequently used for pronouns (*das*, *es*) even when the antecedents are male or female. When the gender of the antecedent was neuter, the gender of the pronoun always matched. If the gender of the antecedent was feminine or masculine, e.g., for "tube" (*der Schlauch*), the gender of the pronoun did not always match. These results can be incorporated in automatic reference resolution of pronouns, and thus enable the resolution also of pronouns where the gender of pronoun and antecedent is not congruent.

In general, a large amount of pronouns and underspecified noun phrases need to be resolved via additional channels. For reference resolution, (i) hands should be continuously tracked and the object at which the hand is directed at needs to be extracted, as well as (ii) the object the instructor grasped last and is currently holding, (iii) the object the instructor grasped before last and is still holding, (iv) information about where a certain object moves or which object or location it touches. This information then needs to be dynamically integrated with the utterance processed.

However, before integrating the information transmitted via these channels and the verbal reference, summaries of tasks at the beginning of task descriptions need to be identified, as well as meta-descriptions referring to the performance of the task. These sentences can be identified via lexical patterns, e.g., "This task is about [...]" (*In dieser Aufgabe geht es darum [...]*).

In addition, general knowledge is needed: (i) knowledge about verbs, (ii) that after connecting two objects of the same type, the connected object can be referred

to via a neuter pronoun, (iii) if nominalised adjectives are referred to one after the other, e.g., "the green (one) and the yellow (one)" (*das Grüne und das Gelbe*) and there is an object containing both attributes, they refer to the same object. For knowledge about verbs, an additional knowledge base is needed including their argument structure, the spatial relation of the manipulated objects before and after the action (e.g., when connecting objects), and whether a hand is involved to conduct the action. The resolution of a neuter pronoun following a "connecting"-action can be implemented via a rule, as well as the resolution of a sequence of pronominalised attributes.

For the resolution of spatial indexicals, visual information is always a prerequisite. However, also information whether there is a NP preceding or following the spatial indexical and if so, whether there is a pause in between is an important cue for reference resolution, as well as whether the already resolved other argument(s) of the verb move to a certain object or location. Including the information that spatial indexicals were only used for somewhere fixed objects (e.g., the mounted tube) or for objects which are part of other objects (e.g., the markers on the tubes) might also enhance automatic reference resolution of spatial indexicals.

In general, robust language processing systems are needed in order to decrease speech recognition errors (Scheutz et al., 2013; Hüwel et al., 2006) and multi-modal information is necessary for the resolution of underspecified verbal references. Therefore, in the task description context, a robot architecture must (i) allow for robust parallel processing of verbal and visual channels, (ii) their temporal alignment, and (iii) seamless integration of information extracted from these channels.

The reliability of the different cues for reference resolution

With regard to reliability, **language** takes a special role: if there is a noun phrase in a situated task description, it is always intended to be a noun phrase, a verb always intended to be a verb, etc., although its lexical content might often not be resolvable via language alone. For verbal object reference, the question thus is not whether it is reliable, but whether it is resolvable to a visually perceivable object. Hence, as a first step, artificial systems need to interpret part-of-speech tags of the incoming utterance as a cue itself in order to identify object references. Subsequently, for reference resolution to objects, the uttered noun phrase can then be compared with the visually perceived object.

Although **gestures** were rarely the only cue to resolve references, they are still very important for directing attention. However, some gestures are also misleading and do not refer to the intended object. These cases can be identified and avoided by

including the following aspects: (i) is the temporal sequence of the gesture so long, that it lasts on during several object references, and (ii) is a person gesturing with two hands in two different directions. The first aspect can be resolved via knowledge about the verb, e.g., its argument structure. Are parallel gestures conducted by the instructor? If yes, is one of the two hands moving? If yes, the moving hand might be the one directed at the relevant object. A challenge for robot architectures is that gestures valuable for reference resolution to objects are not only pointing, but also exhibiting and poising gestures. Thus, a robust gesture recognition system is needed that also allows for the detection of exhibiting and poising gestures.

Knowledge about the verb such as its argument structure (e.g., in "You have to take this tube here" (*Du musst den Schlauch hier nehmen*) versus "You have to insert this tube here." (*Du musst den Schlauch hier einfügen*)) is relevant for the resolution of all verbal referring expressions. This information is not sufficient to resolve referring expressions to objects on its own, but it is very valuable information to distinguish between two or more potential objects identified via other cues.

Information about **the object the instructor grasped last and is currently holding** is a major cue to resolve noun phrases with several potential visual referents. In Task 3, this cue was never misleading. However, this cue is less important to resolve references of pronouns and spatial indexicals.

Information about **the object the instructor grasped before the last one and is still holding** is in particular relevant for the resolution of pronouns. Only in cases where this cue was not present in Task 3, the pronoun could be resolved via the object the instructor grasped last. This cue was also never misleading in Task 3.

In case a NP or a pronoun was uttered preceding or following a spatial indexical, *a pause* can be used to distinguish, whether this spatial indexical refers to the same object as the pronoun or NP, or to another object or location.

Visual information about where a certain object moves or which object or location it touches is important information to resolve NPs and spatial indexicals. In tasks, often two objects are close to or touch each other. This information was also frequently needed and provides reliable information for reference resolution. In case an object moves towards another object (e.g., when a verb for "put" is uttered), first the biggest moving object should be selected (e.g., the tube moves towards the motor block) and when it is then clearer that the marker moves towards a pair of green holding, the marker and the pair of green holdings can be selected.

The interlinkage of verbal referring expressions and additional cues for reference resolution

The results have shown a tight interlinkage between the cues relevant for reference resolution and the linguistic form. Also, the multi-modal channels need to be checked in a certain order to successfully resolve reference. While the first four steps for the resolution of referring expressions overlap, the subsequent – and most relevant cues for the accordant linguistic form – differ.

Noun phrases with several potential visual referents as well as underspecified noun phrases need the following multi-modal cues for reference resolution in the following sequence:

1 Is the utterance a summary of the task before starting the task description in detail? These kind of utterance can be identified via lexical markers and can be ignored for reference resolution to objects in the shared environment. If not:

2 Is the utterance a meta-description, e.g., referring to the performance (e.g., "I am not very good in doing this" (*das kann ich nicht so gut*), "it is a bit difficult" (*es geht ein bisschen schwer*)? These utterances can also be identified via lexical markers and they can also be ignored for reference resolution to objects. If not:

3 Extract information about the verb including their argument structure, the spatial relation of the manipulated objects before and after the action, and whether a hand is involved to conduct the action. Is/are the other argument(s) of the verb already resolved? This information might be needed in an upcoming step.

4 Is a deictic gesture conducted by the instructor? If yes, check the plausibility (according to information extracted about the verb), whether the object gestured at could be the object referred to. If it is plausible, extract the object.

5 Has the instructor grasped an object and is currently holding it? If yes, check the plausibility (according to information extracted about the verb), whether the object gestured at could be the object referred to. If it is plausible, extract the object.

For pronoun resolution, the object the instructor grasped before the last one and is still holding is the most relevant. In case this cue is not present, the object, he/she grasped last can be used for resolution purposes:

1 Is the utterance a summary of the task before starting the task description in detail? If not:

2 Is the utterance a meta-descriptions, e.g., referring to the performance? If not:

3 Extract information about the verb. Is/are the other argument(s) of the verb already resolved? This information might be needed in an upcoming step.

4 Is a deictic gesture conducted by the instructor? If yes, check the plausibility (according to information extracted about the verb), whether the object gestured at could be the object referred to. If it is plausible, extract the object.

5 Has the instructor grasped an object before the last object he/she grasped and is still holding it? If yes, check the plausibility (according to information extracted about the verb), whether this object could be the object referred to. If it is plausible, extract the object.

6 In case the instructor has not grasped an object before grasping the last one, has he/she grasped an object at all and is still holding it? If yes, check the plausibility (according to information extracted about the verb), whether this object could be the object referred to. If it is plausible, extract the object.

Spatial indexicals refer more often to fixed than to loose objects. For the resolution of these verbal referring expressions, the most important cue is thus whether the already resolved argument(s) of the verb move(s) towards an object, or just moved towards and now touch(es) an object:

1 Is the utterance a summary of the task before starting the task description in detail? If not:

2 Is the utterance a meta-descriptions, e.g., referring to the performance? If not:

3 Extract information about the verb. Is/are the other argument(s) of the verb already resolved? This information might be needed in an upcoming step.

4 Is a deictic gesture conducted by the instructor? If yes, check the plausibility (according to information extracted about the verb), whether the object gestured at could be the object referred to. If it is plausible, extract the object.

5 Is there a NP immediately preceding or following the spatial indexical? If yes, in case there is no pause in between, the spatial indexical is probably a reference to the same object as the NP, while it is probably referring to another object or location in case there is a pause in between.

6 Based on information about the verb, do the already resolved arguments of the verb move or did they immediately move before and are now touching an object? If yes, extract the object.

For robot architectures, the parallel processing of verbal as well as non-verbal cues is needed in order to extract and merge the information transmitted via different channels.

The use of temporal markers

Temporal markers communicated by the instructor can be used to structure the task. They can either take the form of verbal markers, such a "first" or "second". Information that the task has come to an end is transmitted via a step back by most of the instructors when the learner only had to observe and listen and only in some cases when the learner was also involved. In Task 4 where the instructor was not involved in the navigation, but only the learner, instructors did not step back. Thus, depending on the task, this information can be used to detect whether the task description has ended. In addition and independent of the task, instructors frequently mentioned at the end that the task was finished, e.g., "That was the task" (*Das war die Aufgabe*). This can also be included and identified via lexical markers.

6.3 Summary and Discussion

In this chapter, challenges and design suggestions for robot architectures were formulated based on the results presented in the previous chapter.

6.3.1 Challenges for human-robot interaction

In Section 6.1, two challenges were presented with regards to the verbal part of referring expressions, and two with regards to additional cues needed for reference resolution.

Variation in wording. The variation of noun phrases when referring to one specific object is very prevalent in the data. Expressions for referring to one single

object varied largely not only between but also within tasks. Different kinds of underspecified noun phrases, such as NPs lexically underspecified for conceptual content, NPs with a similar specific content although not a synonym, and NPs with several potential visual referents need to be resolved.

Pronoun resolution. Also for the resolution of pronouns, meaning negotiation plays a major role. Personal pronouns denoting who conducts the task need to be identified. Also, pronouns referring to objects frequently lack a proximate, congruent and specific antecedent in discourse and the gender of the pronoun often does not match the gender of the antecedent in German. For spatial indexicals such as "here", visual information is a necessary prerequisite for reference resolution.

Verbal and non-verbal cues. Cues necessary to resolve references in situated task descriptions are (i) language, (ii) the objects in the visual field, (iii) gestures, (iv) the object the instructor grasped last and is currently holding, (v) the object the instructor grasped before the last one and is still holding, (vi) knowledge about the verb, (vii) a pause before or after spatial indexicals (viii) visual information about where a certain object moves or which object or location it touches, (ix) the identification of parallel gestures, (x) the identification of a summary of the task at the beginning of the task description, (xi) the identification of a meta-description referring to the performance, (xii) knowledge that after assembling two objects of the same type, they are not referred to separately anymore, and (xiii) knowledge that two colours mentioned one after the other refer to the same object, if there is an object containing the mentioned colours. While the first eight cues are rather general for situated task descriptions, the last two might be specific for the objects and actions occurring in Task 3.

Reliability and interlinkage of different cues. Their reliability as well as their interlinkage still poses a major challenge to robot architectures. Depending on the linguistic form, different cues as well as different sequences of cues in the resolution process are needed.

6.3.2 Lessons for agent design

In Section 6.2, design suggestions were formulated in order to deal with the accordant challenges.

Variation of noun phrases when referring to one specific object. Even if the noun phrase is underspecified and not sufficient for reference resolution, linguistic information might be used to narrow down the number of potential objects. E.g., if "marker" is uttered and there are two different markers, visual cues can then be used to disambiguate between the two objects. Also for nouns or verbs carrying similar semantic information, a strategy to deal with these references would be to search for similar semantic concepts in the visual field. The similar semantic concepts need to be searched for depending on the method of representation.

Pronoun resolution. Personal pronouns to agents can only be interpreted literally, if the learner does not only have to observe and listen, but is actively involved in the task. In all other cases, these personal pronouns in situated task descriptions can be substituted by "I" or "one".

In German, the gender of pronouns is often not congruent with the gender of the antecedent. For automatic reference resolution, the gender of neuter pronouns should also be able to match to a female and masculine antecedent.

Also the use of pronouns without antecedents poses a challenge for robot architectures. In these cases, as well as other cases of pronoun resolution, visual cues need to be incorporated.

The reliability of the different cues for reference resolution. With regard to language, the knowledge about the part-of-speech of the utterances itself is valuable information for a robotic system. This information is rather reliable and can be used as a marker that there is a reference to e.g., an object or an action. In speech, also pauses might occur before or after uttering a spatial indexical and a noun phrase. In case there is a pause between the noun phrase and the spatial indexical, it needs to be interpreted as a distinct object or location reference. In case there is no pause, it can be interpreted as the same object reference as the noun phrase, if information about the argument structure of the verb is not contradicting this hypothesis. Gestures recognised by the system should if possible not only include pointing, but also poising and exhibiting gestures, as they can also be used for reference resolution. While the argument structure of a verb can itself not be used to resolve references, it is often needed to disambiguate between a subset of objects or e.g., to include the other (maybe already resolved) argument of the verb in the resolution process. In this respect, also the object the instructor grasped and is currently holding and the object the participant grasped before the last one and is still holding are relevant. While the first aspect is mainly relevant for the

resolution of a noun phrase referring to several potential objects, the second one is mainly relevant for the resolution of pronouns. Visual information about whether an object moves towards or touches a certain location is mainly relevant for the resolution of spatial indexicals. In order to extract these visual cues, hands of the instructor, as well as objects and their spatial relationship need to be continuously tracked. Depending on the linguistic form of the verbal referring expression, not all non-verbal cues are needed.

The interlinkage of verbal referring expressions and additional cues for reference resolution. Based on the empirical data, a sequence in which the different cues need to be checked was extracted. Depending on the linguistic form, the relevance of the various cues differs. The first four steps, however, are the same for all linguistic references: (1) It needs to be identified whether the utterance is a summary of the task description or (2) a meta-description. (3) Information about the verb is then extracted as it might be needed in addition to visual cues. Subsequently, (4) the visual field is checked for gestures.

For underspecified noun phrases, the major visual cue is if (5a) the instructor has just grabbed an object and is currently holding it.

For the resolution of pronouns, it is important whether (5b) the instructor has grasped an object before the last one and is still holding it. Only if this cue is not present, information about (6b) the object the instructor just grasped and is currently holding can be used for reference resolution.

For spatial indexicals it is relevant whether (5c) there is a NP and a pause preceding or following the spatial indexical and whether the pause is between the noun and the spatial indexical or not. In case it is in between, the two verbal referring expressions are not directed at the same object. Subsequently, it needs to be extracted whether (6c) the already resolved arguments of the verb move or did immediately before move and are now touching an object. This object then needs to be extracted.

A subset of the above presented design suggestions for robot architectures are already implemented in a reference resolution mechanisms. This collaborate work with Tom Williams, Saurav Acharya, and Matthias Scheutz will be presented in the next section.

Chapter 7

Preliminary Implementation

The following chapter presents collaborate work with Tom Williams, Saurav Acharya, and Matthias Scheutz from Tufts University about the development of an algorithm for situated open world reference resolution. In Section 7.1, work related to computational models of the Givenness Hierarchy by Gundel et al. (1993) is discussed. Based on the empirical results (presented in Chapter 5.3) we adapted a version of the Givenness Hierarchy (Section 7.2) and the colleagues developed an algorithm for open world reference resolution (Section 7.3). Subsequently, we validated and evaluated the algorithm on a translated subset of the data presented in this thesis (Section 7.4) and the chapter will conclude with a summary and a discussion in Section 7.5.

This work is in line with the focus on object reference resolution in multi-modal task descriptions and I contributed to the theoretical part of adapting the Givenness Hierarchy to the empirical data and I provided the data and general input for the evaluation of the algorithm. In the following, the main aspects of this collaborate work will be summarized. For more details, see Williams et al. (2015) and Williams et al. (2016).

7.1 Background

Motivation for this work is the need for robots interacting with humans to create or resolve references to given or new entities in natural language dialogue similar to the objectives of this thesis. In the following, an algorithm is proposed for resolving references of the following linguistic forms: definite and indefinite noun phrases, as well as pronouns. Spatial indexicals are not taken into account in the current version. Basis for this approach is the Givenness Hierarchy by Gundel et al. (1993).

The Givenness Hierarchy contains six levels of "cognitive accessibility": in focus

$\subset$ activated $\subset$ familiar $\subset$ uniquely identifiable $\subset$ referential $\subset$ type identifiable. The hierarchy is nested, so any information that is *in focus*, is also *activated, familiar* etc., any information, that is *activated*, is also *familiar, uniquely identifiable* etc. but not *in focus* and so on. The Givenness Hierarchy together with its "coding protocol" (see Gundel et al., 2006) provide a solid framework for reference resolution, including (i) data structures needed for reference resolution, (ii) guidelines for how to populate these data structures, and (iii) guidelines for how to retrieve information from those data structures. Additionally, the framework is based on empirical findings (see Gundel et al., 2006).

In human-robot interaction, several efforts have made use of the Givenness Hierarchy. To the best of our knowledge, there exist only two other implementations that made full use of the Givenness Hierarchy, see also Chapter 3.1, page 27 for a short summary (for a better readability, the approaches are here again briefly summarized). In the first approach by Kehler (2000), the levels "referential" and "type identifiable" were omitted, because they were mostly interested in pen-and-tablet interfaces and in this context, unknown or hypothetical references are unlikely. Kehler's approach is based on four rules: (i) if an object is gestured to, choose that object; (ii) otherwise, if the currently selected object meets all semantic type constraints imposed by the referring expression, choose that object; (iii) otherwise, if there is a visible object that is semantically compatible, then choose that object; (iv) otherwise, a full NP was used that uniquely identified the referent.

Chai et al. (2006) noticed two drawback in the algorithm by Kehler (2000): (i) employing the four rules, it is impossible to identify or resolve ambiguities, and (ii) with this implementation, multiple references in one utterance cannot be handled.

Chai et al. (2006) implemented a greedy algorithm for combining the Givenness Hierarchy with Conversational Implicature by H. Grice (1975) which is able to handle ambiguities and multiple references in one utterance. Their modified hierarchy looks the following: *gesture* $\subset$ *focus* (subsuming Gundel's *in focus* and *activated* tiers) $\subset$ *visible* (subsuming Gundel's *activated* and *uniquely identifiable* tiers) $\subset$ *others* (subsuming Gundel's *referential* and *type identifiable* tiers).

We (see Williams et al., 2015) argue that this adaptation of the hierarchy and its implementation are insufficient for realistic human-robot interaction scenarios, due to the following reasons: (i) in human-robot interaction the robot does not always know with complete certainty whether or not an entity has a certain property (ii) multiple referring expressions within one utterance to entities not currently visible need to be handled, and (iii) references to events, speech acts, or entities that cannot physically exist need to be resolved, (iv) the levels *in focus* and *activated* need to be

differentiated in order to be able to distinguish between *Can you repeat it?* and *Can you repeat that?*, and (v) subsequent referential expressions need to be resolved, if an incorrect referent is chosen for the first one (which is not possible using a greedy approach).

Despite the modifications to the implementation, we used the results of Section 5.3 to suggest modifications to the Givenness Hierarchy itself.

7.2 Adaptation of the Givenness Hierarchy based on Empirical Results

We adapted two areas of the Givenness Hierarchy based on the results: (i) gaze and gesture handling; as well as (ii) information retrieval. In Task 2 as well as Task 3, participants frequently uttered underspecified definite noun phrases as verbal referring expressions to objects they were currently looking at. This is consistent with the coding protocol of the Givenness Hierarchy, in which entities that are the subject of speech-simultaneous gesture or gaze should be considered to be "activated". However, it is not always possible to identify the entity at which an interlocutor is looking. Therefore, we suggest that all entities in the interlocutor's field of view are salient and thus "activated". In Gundel et al. (2010) they suggest that there are different degrees of being "in focus". We suggest to differing degrees also for entities to be "activated". Thus, all entities in the interlocutor's view could have different activation scores, depending on how recently the interlocutors looked at them.

In case ambiguous nouns are uttered, e.g., "tube", or "marker", and there is more than one of these entities, their activation scores could be increased when the target entity is gazed or pointed at. If the entities within a particular level are ordered in decreasing order of activation, the entity referred at will naturally be arrived first. Thus, entities can be resolved without the explicit use of gaze or gesture checking.

Underspecified noun phrases, such as "thing" (*Ding*) or "object" (*Objekt*) occurred in both tasks. These kind of references could also be resolved employing the suggested approach.

Linguistic form cueing entities "in focus" could often not be resolved in the data via automatic anaphora resolution, either due to the distance between the form used and the last reference to the entity, or a lack of congruence, e.g., the gender of the pronoun does not match the gender of the antecedent. Thus, we suggest to use gesture and eye gaze to increase not only the activation level, but also the focus level. This is in contrast to approaches presented by Gundel et al. and Chai et al., as neither use visual cues when considering entities on the "focus" level.

Table 7.1: The suggested search plans.

Level	**Search plan**
"focus"	FOC
"activated"	ACT –> FOC
"familiar"	ACT –> FOC –> FAM
"definite"	ACT –> FOC –> LTM
"this N activated"	ACT –> FOC –> HYP
"indefinite"	HYP

In the data, instructors often referred to the board as "the board", which would cue the "uniquely identifiable" level. According to the Givenness Hierarchy, the board in the task setting is a less appropriate choice than any other board the listener has previously encountered. The same accounts for the "familiar" level. We thus suggest to consider the "activated" and "in focus" level before considering the levels the referential forms actually cue.

While it is difficult to differentiate whether *this N* cues the "activated" or the "referential" level, we believe that a first step would be to treat all uses of *this N* as "activated" so long as a suitable entity can be found at the "activated" or "focus" level, see Table 7.1 for a summary of the search plans.

7.3 The Algorithm

First, **parsing and analysis** of the utterances will be discussed. Subsequently, the data structures will be described and how they are used to resolve references in parsed utterances. As a framework, the Distributed, Integrated, Affected, Reflection and Cognition (DIARC) architecture is used (Scheutz et al., 2007), as implemented in the Agent Development Environment (ADE) (Scheutz, 2006; Scheutz et al., 2013).

Subsequently to sending each utterance to the C&C parser (S. Clark & Curran, 2007), a dependency graph is extracted, converted into a tree and the following information extracted: (i) a set of formulae representing the surface semantics of the utterance, (ii) a set of "status cue" mappings for each referenced entity, and (iii) the type of utterance, which was heard (e.g., "statement").

For **data structure population**, the data structures *FOC*, *ACT*, *FAM*, and *LTM* are used, corresponding with the first four levels of the Givenness Hierarchy. A summary of how the data structures are populated are presented in Table 7.1. FOC and ACT are reset after each clause and FAM after each dialogue. The levels (despite

Table 7.2: The suggested search plans. Lines marked with an asterisk represent future work.

Level	Contents
FOC	Main clause subject of clause n-1
	Syntactic focus of clause n-1
	* event denoted by clause n-1
ACT	* Entities visible in the int's region of attention
	all other entities referenced in clause n-1
	* Focus of int.'s gesture, if any
	* Focus of int.'s sustained eye gaze, if any
	* Speech act associated with clause n-1
	* All propositions entailed by clause n-1
FAM	All entities referenced in clause n-1
	* The robot's current location
LTM	All declarative memory

LTM) are then updated using the rules listed in Table 7.1. The syntactic focus and event denoted by clause n-1 are placed in FOC, the speech act and any propositions entailed by clause n-1 are placed into ATC, and all entities referenced at all in clause n-1 are placed into both ACT and FAM. Additionally, locations visited by the robot are placed into FAM and entities in the interlocutor's region of attention into ACT. Within each data structure, the entities are then sorted depending on whether they are in the main clause ($m(e) \in [0, 1]$) , syntactic prominence ($s(e)$), and their recency of mention ($r(e)$). As a scoring function, $\Gamma(e) = \alpha_1 * m(e) + \alpha_2 * s(e) + \alpha_3 * r(e)$ is used, where α_1, α_2, and α_3 are monotonically decreasing coefficients prioritizing the three measures. Extra-linguistic factors are not included yet and part of future work.

For **reference resolution**, three algorithms are employed. The first one (GH-POWER) collects the variables appearing in S (the semantics of clause n), and sorts them with respect to the level they are cued towards. Additionally, the module POWER, a module for Probabilistic, Open-World Entity Resolution, to interface with LTM (as described in Williams & Scheutz (2015)), is employed. If for example X is cued towards "in focus" and Y towards "familiar" in M (the status cue mappings for clause n), then X will appear before Y. GH-POWER then initiates a cache-table which stores a memorized list of variables and levels, i.e., FOC, ACT, FAM and

HYP. As a next step, the data structures need to be determined where to look for those entities, based on the search plan in Table 7.2. All multi-variable plan combinations are stored and an empty set of candidate hypotheses is created. GH-POWER first separates variables for which the LTM must be queried from all other variables. Next, it iterates over each pair in the row. For example, if there is no hypothesis yet about the first entry in this row (e.g., *Y* and ACT), and if it is not HYP, GH-POWER uses the algorithm ASSESS to search the according level for the most likely entity. It creates for each entity of the according level a new hypothesis which maps the entity to the variable of interest. For example, if the sentence "The ball in this red box." is heard and there is one entity in ATC (e.g., *obj_1*), the algorithm would consult the module POWER to see to what degree *obj_1* could be considered to be a box and to what degree it could be considered to be red, and then creates a hypothesis mapping Y to obj_1. Once, all formulae containing only variables (e.g., Y) are examined, all those containing both variables and any other previously examined variables are examined using the third algorithm: ASSESS-ALL. If a sentence contains two references (e.g., a ball and a red box), ASSESS-ALL would inquire to what degree the candidate entities for X could be considered to be "in" each candidate entity for Y (e.g., a ball is in a red box). For example, the probabilities that X is obj_2 (0.75) and Y is obj_1 (0.9) these probabilities are combined, as well as the probabilities that X is obj_3 (0.9) and Y is obj_1 (0.9). If ASSESS determines that *in*(X;Y) has probability 0.9 for the first hypothesis and 0.1 for the second, the two hypotheses are updated and obj_2 is selected for X and obj_1 for Y.

Subsequently, GH-POWER considers all variables set aside to be searched for in LTM and each candidate binding in a set of candidate hypotheses. For each variable, a set of formulae is bound using the variable binding of the accordant candidate hypothesis and an ordering of the variables to be queried in LTM is created based on the prepositional attachment observed in the set of formulae. The POWER algorithm is then used to determine (i) whether the variables refer to unknown entities, and (ii) which entities in the LTM are the most probable referents for each other variable.

Finally, the number of remaining hypotheses is examined. If more than one or no hypothesis was found, GH-POWER returns a set of solutions, i.e., the expressions are either ambiguous or unresolvable. If only one hypothesis remains, GH-POWER is used to update the set of semantics and POWER to assert new representations for each variable. For example, in an utterance X is bound to obj_4 and Y to "?", and a single hypothesis is produced with probability 0.7. POWER will then be used to

create a new object with properties. Once all partitions have been precessed, the results are combined into a comprehensive set of candidate binding hypotheses.

7.4 Validation and Evaluation

In this section, an evaluation of the proposed algorithm on the human-human and human-robot interaction data of Task 3 is presented. The algorithm was provided with a knowledge base containing information about the robot's environmental and task context and then incrementally processed the utterances. First, several test cases were evaluated to demonstrate the success of GH-POWER with regard to our concerns with previous approaches based on the Givenness Hierarchy. We confirmed the follow aspects, that could not be handled by previous approaches:

Uncertainty: When the robot believed two entities (e.g., tubes) are rated to be a certain object (e.g., the flexible tube), the entity with the highest probability is chosen by GH-POWER.

Open worlds: When a robot only knew a red and yellow marker, GH-POWER posited a new entity when resolving "Find the blue marker".

Hypothetical entities: When the robot knew of a box on a table and was asked to "Imagine a box. Describe the box.", "the box" was resolved to the imaginary box and not to the one on the table.

Unobservable entities: When the robot believed it was learning a task and was asked to "Describe the task.", GH-POWER correctly resolved "the task".

Complex noun phrases: The utterance "Pick up the tube that is on the triangular table." could be correctly resolved by GH-POWER, when there was a "familiar" tube on a triangular table and an "activated" tube on a round table.

Evaluation. To evaluate, the algorithm, the participants' utterances of Task 3 were translated to English and disfluencies as well as parenthetical statements were removed. A knowledge base of objects and agents was constructed and provided to GH-POWER. Then, each task-relevant utterance was provided to GH-POWER and compared to the annotations. The algorithm identified 270 references (not only to objects, but all references) in the HH dyads and 98 in the HR dyads. However, 17.93% of the references (19.26% in the HH dyads and 14.29% in HR dyads) found by the C&C parser were not references but artefacts or parse errors. Discarding

these parse errors, GH-POWER correctly resolved 55.50% in HH dyads and 57.14% in HR dyads.

The human-human interaction contained 110 task relevant utterances, the human-robot interaction 32. GH-POWER resolved 44.81% (121 out of 270) in the HH dyads and 48.98% (48 out of 98) references in the HR dyads. The other references could not be resolved due to the following reasons:

Plurals (5.96% in HHD, 2.38% in HRD): GH-POWER is currently unable to handle plural references and this will be part of future work.

Non-discrete entities (10.55% in HHD, 10.71% in HRD): References to resolve regions or sections of the tube could not be handled and will be part of future work.

Gestural information (10.09% in HHD, 10.71% in HRD): While it is an explicit design aim to handle gestures, they as well as eye gaze are not included yet and will be part of future work.

Summarisation at the beginning of the task (5.96% in HHD, 1.19% in HRD): References made at the beginning of the task were difficult to resolve, because speakers shared a joint context and additional knowledge was needed, e.g., in "I will now describe *it* to you".

Idiomatic or colloquial references (4.13% in HHD, 1.19% in HRD): To be able to resolve references such as "That was it." a tighter integration of GH-POWER with pragmatic inferences may be needed.

Low linguistic salience scores (1.32% in HHD, 1.19% in HRD): Some references were incorrectly resolved because the linguistic salience score were not sufficient to select the target. Thus, other salience functions will be needed to investigate in future work.

Various other reasons (6.42% in HHD, 15.48% in HRD): For example some instructors referred to concepts we were unprepared to handle (e.g., "The problem here is..."), or they used noun phrases in ways we did not anticipate (e.g., "There is a pipe there.").

7.5 Summary and Discussion

In collaboration with colleagues at Tufts University, we have proposed, validated and evaluated a reference resolution algorithm, suitable for resolving the majority of references naturally occurring in situated, task-based interactions. GH-POWER uses an adapted version of the Givenness Hierarchy, allowing additionally for (i) inter-traversal, (ii) salience-based intra-tier candidate selection, and (iii) multiple

resolution. It is thus able to handle a wider range of referring expressions than previous reference resolution algorithms employing the Givenness Hierarchy. Additionally, unlike previous approaches, open world and uncertain contexts can be handled and it is thus better suited for human-robot interaction than previous algorithms employing the Givenness Hierarchy.

Our contribution to the state of the art in human-robot interaction are the following: First, we use the complete Givenness Hierarchy to handle linguistic forms and resolve references which were not covered by previous implementations of the Givenness Hierarchy. Second, the proposed algorithm is able to handle uncertain and open worlds. Third, the algorithm includes findings from human-human situated task-descriptions and is thus empirically verified. Finally, the algorithm is thus a good starting point for studying the interaction of cognitive processes, both for artificial agents as well as cognitive modelling research.

Future work with regards to the algorithm includes the following aspects:

- Data structures need to be populated also with entities observed visually, gestured at, and other non-verbal cues.
- How to determine whether a referential expression is cueing "familiar" or "referential" needs to be further investigated.
- The best way to calculate activation and focus scores needs to be determined.
- Plural and non-discrete references were relatively common in task-based dialogues. Dealing with these categories needs to be further investigated.
- Common-sense reasoning capabilities need to be integrated, e.g., to enable the robot to determine that some referents are less likely given, e.g., the action it is being asked to perform on that referent.
- If a set of resolution hypotheses is produced, the robot must alert its interlocutor and generate clarification requests.

Chapter 8

Conclusion

The goal of this thesis was to extract general principles of multi-modal human-human task descriptions, identify challenges for robot architectures and formulate design suggestions for robot architectures in order to tackle the identified challenges. The rationale behind this objective was to shed light on the interplay of verbal and non-verbal cues in situated task descriptions, identify aspects causing variation as well as cues that enable a learner of situated task descriptions to still be able to identify, extract, and merge all information necessary for the task.

As a starting point, general aspects of human situated task descriptions were collected in the literature which might pose challenges to situated human-robot task descriptions. They included general inter- and intra-speaker variation with regard to language, verbal referring expressions to objects, as well as multi-modal cues. Existing theoretical as well as computational approaches were reviewed and discussed. Although some theoretical and computational models include pointing gestures, eye gaze, or objects in the visual field as non-verbal cues, they stay either very vague on how these cues need to be interlinked with language, or the non-verbal cues are more of an add-on to language. However, these approaches are not sufficient to account for situated, task-based descriptions.

A pilot study was analysed to gain first insights in the variation of how instructors transmit one and the same task. Additionally, the applicability of a cognitive model for human embodied language comprehension was tested on the empirical data. Based on first results on situated task descriptions and the model for embodied language comprehension, an annotation framework was developed for situated task descriptions.

Based on the results of the pilot study, a setup for an empirical human-human and human-robot data collection was designed and data suitable for investigating situated task descriptions were collected. An analysis of the data has shown that in

situated task descriptions, language takes more of a secondary role for information transmission. From language, it can be determined whether there is an object reference or an action, e.g., via information about part-of-speech. Although eye gaze of the instructor was directed at the object referred to, it was directed somewhere else in the other half of the cases. Thus, it is not a very reliable cue for automatic references resolution. However, gestures and other non-verbal cues including the position of the hands of the instructor and the relation between objects needs to be continuously tracked. Based on knowledge about the verb, this information then needs to be merged in a certain sequence. Also, the consulted cues depend on the linguistic form of the verbal referring expression.

As an upshot, the relevant linguistic and visual information will need to be incrementally incorporated in a robot architecture for the robot to be able to resolve referents. The resulting design suggestions have high potential to enhance human-robot situated task-based interaction. Some resulting design suggestions have already been implemented in collaborate work in an algorithm for open-world reference resolution and also tested on a subset of the data.

In the following, implications for both research on human-human interaction, as well as research on artificial agents are outlined. Subsequently, limitations of the work are discussed with potential for future work.

8.1 Implications

Due to the widely open research question which variation occurs in human-human situated task descriptions with focus on reference resolution to objects, the results of this work provided first steps towards a more thorough understanding of information extraction in human situated task descriptions. These results served then as a basis for formulating design suggestions in order to enable robots to be able to extract information necessary for the task. Implications for both research areas will be summarised in the following.

8.1.1 Implications for research on human-human interaction

In order to investigate human-human interaction data from the point of view of what the robot is exposed to is also valuable for research on human situated task-based interaction because from this viewpoint, all cues potentially transmitting information need to be incorporated in the analysis. The empirical data revealed that language takes more of a scaffolding role in situated task descriptions. Via language it can be identified, when, for example, an object reference is uttered. However, the portion

of references that can be resolved via language is very small. This calls in general for a tight interlinkage of verbal and non-verbal cues. In theoretical models, e.g., for reference resolution, some include non-verbal cues. However, they include at most eye gaze, gestures, and objects in the visual field. The empirical analysis of the data presented in this thesis has shown that there are more cues relevant for object reference resolution to objects in situated task descriptions: (i) verbal referring expressions which can either directly be resolved to an object in the visual scene or which need to be combined with information transmitted via other channels, (ii) pointing, poising, and exhibiting gestures, (iii) the object the instructor grasped last and is currently holding, (iv) the object the instructor grasped before the last one and is still holding, (v) visual information about where a certain object moves or which object or location it touches. In addition to these cues, knowledge about the verb such as argument structure is potentially needed for reference resolution.

This thesis also provided detailed information on the interlinkage between linguistic form and non-verbal cues, as well as the sequence needed for reference resolution. These aspects necessary for reference resolution of situated task descriptions have neither been considered in current theoretical/psycholinguistic nor in computational models for references resolution.

In addition, a cognitive model of embodied language understanding (see Zwaan, 2004) has been employed on empirical, situated task descriptions. Although this model includes important aspect of human cognition and language comprehension based on action and perceptions, the results have shown that it is not suitable for situated task descriptions. It lacks many important aspects of situated task descriptions, such as a vast variation in wording and information not transmitted via language but via other cues. On the other hand, sharing representations of concepts with the interlocutor is not as important in situated task descriptions as it is in interactions not focussing on objects or actions in the shared environment.

To summarise, most approaches and models do not deal with situated task descriptions and if they do, they only include non-verbal cues on a level too rudimentary for references resolution. They lack details on the interplay of different modalities and they are thus not sufficient for implementation. However, in human-robot interaction, situated task descriptions are very frequent and insights from human-human interaction are a necessary prerequisite to develop mechanisms for robot architectures to deal with this kind of communication context.

8.1.2 Implications for research on artificial agents

From the point of view of research on artificial agents, the results showed the need for a tight combination of verbal and non-verbal information processing in situated task descriptions. While some current models of reference resolution account for eye gaze of the instructor, gesture, as well as objects in the visual field of the interlocutors, according to the results presented in this thesis, eye gaze is often not directed at the referred object and it is thus not reliable enough to be used as a cue by robot architectures. This is not to say that this thesis means to devalue eye gaze as a referential cue. In about half of the cases participants looked at the object they intended to refer. Still, in the other cases, they looked at the object or location they last manipulated or they planned to manipulate in the future. Also, they looked back and forth between two objects they had to mount etc. Therefore, for the automatic extraction of objects referred to, other visual cues are more reliable. The thesis provided a comprehensive list of linguistic and visual information needed, in order to resolve references to objects, as well as the interlinkage of cues and the sequence, in which they need to be searches for to resolve references to objects.

A subset of general principles has already been used to adapt a version of the Givenness Hierarchy by Gundel et al. (2006) in order to deal with situated, task-based references resolution.

Besides cues for reference resolution, general lessons for agent design were proposed to enhance situated human-robot task descriptions, as the resolution of personal pronouns referring to agents, or references to actions. The results highlight the tight interlinkage between the discourse model and the interaction model. When focussing solely on discourse, reference resolution is doomed to fail.

In the following, the major outcomes of this thesis with potential for enhancing situated human-robot interaction are summarized:

> The most important outcome of this thesis are design suggestions for robot architectures in order to enhance human-robot interaction in situated task descriptions. They are formulated on the basis of three different analyses and do not only include but focus on the resolution of object references.
>
> The annotated data collected have potential to serve as empirical basis for various future research questions in the context of situated task descriptions. They were collected in such a way to allow many different approaches to the data material, e.g., including comprehensive recordings of video, audio, motion, and force data.

Results of the pilot study have shown that general models for embodied language comprehension are not adequate to investigate situated task descriptions, as they focus on internal representations and processes during language comprehension if the objects referred to are not in the shared environment. However, in Chapter 4 of this thesis, an annotation scheme was developed which combines a model for embodied language comprehension (the Immersed Experiencer Framework) with properties of situated task descriptions. This annotation framework or parts of this framework (depending on the research questions) can be of value for future analysis of situated task descriptions.

The version of the Givenness Hierarchy adapted on the empirical results presented in this thesis as well as the validated and evaluated algorithm for open world reference resolution has high potential to enhance human-robot interaction.

8.2 Outlook

The approach presented in this thesis goes beyond related work in human-human and human-robot multi-modal situated task descriptions. Nevertheless, there is, of course, room for future research:

Bias of introducing instructors by words. The experimental design has the bias of introducing the future instructor using language. If no variation would have been found, the bias would have been a problem. However, the bias was not strong enough, as there was still a wide range of variation in word choices. If the future instructor were not introduced verbally, at least the variation could be expected that was found in the presented data. Still, this could be addressed in future work.

Communication is a bilateral process. Instructors' and learners' actions are likely to be coupled (H. H. Clark & Wilkes-Gibbs, 1986). However, the main research of this thesis focuses on which channels instructors transmit information through. Therefore, the focus in this thesis is on the instructors only. However, learners' activities will need to be addressed in future work.

German data. The data were collected in German by native speakers. Gestures, for example, are widely observed behaviours in human interaction across various contexts and cultures (see Argyle & Cook, 1976; McNeill, 1992). However, it is still possible that in a different cultural environment similar ex-

periments lead to different results. Also, communicative behaviour can vary between an experimental setting and the corresponding real-world situation the experiment seeks to emulate, or according to the relationship between instructor and learner. The complexity regarding how humans index objects by means of lexical choice, eye gaze, and gesture and other non-verbal cues is difficult to do justice to. The data presented in this thesis already show a wide range of variation, although the experiments were conducted in a laboratory setting and the future instructor was acquainted with the tasks verbally.

Further analysis of the data material. Due to the research questions presented in this thesis, various aspects of the comprehensive data collection will be left to future research, e.g., temporal sequences of the different cues within the referring expressions to objects, referring expressions to actions, the role of force in collaboratively manipulating an object, the role of prosody in referring expressions, and others.

To conclude, results of the recorded and annotated data seem to be suited to define challenges and formulate design principles for robot architectures to deal with situated task-based instructions. Thus, the work presented here can also serve as a basis to further investigate the collected data.

Zusammenfassung

Ein Roboter muss mit einer großen Variation verbaler und non-verbaler Information umgehen können, um in der Lage zu sein, in situierten Aufgabenbeschreibungen Referenzen auf Objekte, Personen oder Aktionen auflösen zu können. Um Roboter zu entwickeln, die in der Lange sind mit Menschen auf natürliche Art und Weise zu interagieren, muss noch eine Anzahl von Bereichen menschlicher Aufgabenbeschreibungen weiter untersucht werden.

Im Rahmen der Dissertation wurden Daten gesammelt, in denen eine Person jemandem vier kurze Aufgaben erklärt, um menschliche Interaktion in größerem Detailliertheitsgrad untersuchen zu können. Die Analyse dieser Daten ist eine wichtige Basis dafür, womit ein Roboter umgehen können muss, wenn er an Stelle der lernenden Person wäre.

Die qualitative Analyse der Daten hat gezeigt, dass multi-modale Kommunikation bei situierten Aufgabenbeschreibungen eine sehr wichtige Rolle spielt. Wenn nur der sprachliche Teil der Instruktionen interpretiert wird, geht wichtige Information verloren, die notwendig ist, um die Aufgabe erfolgreich durchführen zu können. Neben Augenbewegungen und Gesten ist z.B. auch wichtig, welche Objekte die instruierende Person in der Hand hält, die Argumentstruktur von Verben, oder ob sich die Hand der lehrenden Person gerade zu einem Objekt hinbewegt. Die Relevanz der jeweiligen non-verbalen Beobachtung hängt mit der geäußerten linguistischen Form zusammen. Bei geäußerten Nominalphrasen ist es wichtig, welches Objekt die instruierende Person gerade gegriffen hat, während es bei Pronomina relevanter ist, welches Objekt vor längerem gegriffen wurde, aber immer noch gehalten wird.

Basierend auf den Ergebnissen der Datenanalyse werden generelle Prinzipien formuliert, wie Referenzen in situierten multi-modalen Aufgabenbeschreibungen aufgelöst werden können. Ebenso werden Anforderungen diskutiert und daraus resultierende Design Ideen für Roboter- Architekturen im Bezug auf den Umgang mit (i) einer großen Variation an verbalen Äußerungen, wenn auf ein spezifisches Objekt verwiesen wird, (ii) unterspezifizierten sprachlichen Referenzen, und (iii) ihrer Multi-modalität.

Abstract

A robot has to deal with a broad variety of verbal and non-verbal information to be able to resolve references in a situated task description context. If robots are to interact with humans in the future, a number of issues in natural situated task descriptions need to be tackled.

In order to investigate human-human interaction in more detail, data were collected where an instructor explains and shows four different tasks to a learner. The results are an important basis for what a robot would have to deal with if it were in the learner's position.

The qualitative analysis of the data shows that multi-modal communication plays a crucial role in situated task descriptions. If only the verbal part of task descriptions is used for interpretation, important information for successfully conducting the task is missing. In addition to eye gaze and gesture of the instructor, additional cues are needed for multi-modal reference resolution. These include which object the instructor is holding or still holding, knowledge about the argument structure of verbs, or whether the hand is moving towards an object. The relevance of these cues depends on the uttered linguistic form. For example the object which the instructor grasps at a certain point in time is important when a noun phrase is uttered, while for pronouns the object which the instructor is still holding is more relevant.

Based on the results of the data analysis, general principles of human multi-modal task descriptions are formulated on how references to objects can be resolved and the accordant challenges for robot architectures are discussed. These challenges include (i) a broad variation of verbal referring expressions when referring to one specific object, (ii) verbally underspecified referring expressions and (iii) their multi-modality.

References

Abuczki, Á., & Ghazaleh, E. B. (2013). An overview of multimodal corpora, annotation tools and schemes. *Argumentum*, *9*, 86–98.

Admoni, H., Datsikas, C., & Scassellati, B. (2014). Speech and gaze conflicts in collaborative human-robot interactions. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014)*.

Allen, J., & Core, M. (1997). *Draft of DAMSL: Dialog act markup in several layers* (Tech. Rep.). University of Rochester: Department of Computer Science. Retrieved from `http://www.cs.rochester.edu/research/speech/damsl/RevisedManual/`

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, *41*(3-4), 273–287.

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review*, *106*(4), 748.

Anastasiou, D. (2012). A speech and gesture spatial corpus in assisted living. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)* (pp. 2351–2354).

Andersen, E. S., Dunlea, A., & Kekelis, L. S. (1984). Blind children's language: Resolving some differences. *Journal of child language*, *11*(03), 645–664.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., ... others (1991). The HCRC map task corpus. *Language and speech*, *34*(4), 351–366.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze.* Cambridge U Press.

Ariel, M. (2001). Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, *8*, 29–87.

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., & Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, *76*(1), B13–B26.

Arnold, J. E., & Tanenhaus, M. K. (2011). Disfluency effects in comprehension: How new information can become accessible. *The Processing and Acquisition of Reference*, 197–217.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*(1), 361–374.

Barr, D. J., & Keysar, B. (2006). Perspective taking and the coordination of meaning in language use. *Handbook of psycholinguistics*, *2*, 901–938.

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, *33*(1), 5–22.

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science*, *2*(4), 716–724.

Beattie, G. W., & Butterworth, B. (1979). Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, *22*(3), 201–211.

Bergmann, K., & Kopp, S. (2012). Gestural alignment in natural dialogue. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Berkenfield, C. (2001). The role of frequency in the realization of English that. *Frequency and the Emergence of Linguistic Structure*, *45*, 281-307.

Bock, K., Irwin, D. E., Davidson, D. J., & Levelt, W. J. (2003). Minding the clock. *Journal of Memory and Language*, *48*(4), 653–685.

Böckler, A., Knoblich, G., & Sebanz, N. (2011). Observing shared attention modulates gaze following. *Cognition*, 292–298.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, *5*(9/10), 341–345.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, *44*(2), 123–147.

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, 41–44.

Brennan, S. E. (2000). Processes that shape conversation and their implications for computational linguistics. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (pp. 1–11).

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., & Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, *106*(3), 1465–1477.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493.

Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop Towards a Standard Markup Language for Embodied Dialogue Acts* (pp. 13–24).

Büring, D. (2012). Focus and intonation. In G. Russell & D. G. Fara (Eds.), *Routledge companion to the philosophy of language.* Routledge.

Cantrell, R., Scheutz, M., Schermerhorn, P., & Wu, X. (2010). Robust spoken instruction understanding for HRI. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 275–282).

Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J. C., & Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, *23*(1), 13–31.

Chafe, W. L., & Li, C. N. (1976). Subject and topic. In C. Li (Ed.), (chap. Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View in Subject and Topic). New York: Academic Press.

Chai, J. Y., Prasov, Z., & Qu, S. (2006). Cognitive principles in robust multimodal interpretation. *Journal of Artificial Intelligence Research (JAIR)*, *27*, 55–83.

Chai, J. Y., She, L., Fang, R., Ottarson, S., Littley, C., Liu, C., & Hanson, K. (2014). Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 9th International Conference on Human-Robot Interaction (HRI)* (pp. 33–40).

Ciocarlie, M., Hsiao, K., Jones, E. G., Chitta, S., Rusu, R. B., & Şucan, I. A. (2014). Towards reliable grasping and manipulation in household environments. In *Experimental robotics* (pp. 241–252).

Clark, H. H. (1996). *Using language* (Vol. 1996). Cambridge University Press Cambridge.

Clark, H. H. (2003). Pointing and placing. *Pointing: Where language, culture, and cognition meet*, 243–268.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *Perspectives on Socially Shared Cognition*, *13*(1991), 127–149.

Clark, H. H., & Krych, M. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*(1), 62–81.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Clark, S., & Curran, J. R. (2007). Wide-coverage efficient statistical parsing with ccg and log-linear models. *Computational Linguistics*, *33*(4), 493–552.

Coradeschi, S., Loutfi, A., & Wrede, B. (2013). A short review of symbol grounding in robotic and intelligent systems. *KI-Künstliche Intelligenz*, *27*(2), 129–136.

Crawley, R. A., Stevenson, R. J., & Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, *19*(4), 245–264.

Dael, N., Mortillaro, M., & Scherer, K. R. (2012). The body action and posture coding system (BAP): Development and reliability. *Journal of Nonverbal Behavior*, *36*(2), 97–121.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, *47*(2), 292–314.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*(2), 233–263.

Dautenhahn, K., Walters, M., Woods, S., Koay, K. L., Nehaniv, C. L., Sisbot, A., ... Siméon, T. (2006). How may i serve you?: a robot companion approaching a seated person in a helping context. In *Proceedings of the 1st Conference on Human-Robot Interaction (HRI)* (pp. 172–179).

Dzifcak, J., Scheutz, M., Baral, C., & Schermerhorn, P. (2009). What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Proceedings of the International Conference on Robotics and Automation* (pp. 4163–4168).

Eberhard, K. M., Nicholson, H., Kübler, S., Gundersen, S., & Scheutz, M. (2010). The Indiana" cooperative remote search task"(CReST) corpus. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC).*

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*(6), 409–436.

Eisler, F. G. (1968). *Psycholinguistics: Experiments in spontaneous speech.* Academic Press New York.

Ekman, P., & Friesen, W. V. (1981). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Nonverbal Communication, Interaction, and Gesture*, 57–106.

Fang, R., Doering, M., & Chai, J. Y. (2015). Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the 10th International Conference on Human-Robot Interaction (HRI)* (pp. 271–278).

Fischinger, D., Einramhof, P., Papoutsakis, K., Wohlkinger, W., Mayer, P., Panek, P., ... others (2016). Hobbit, a care robot supporting independent living at home: First prototype and lessons learned. *Robotics and Autonomous Systems*, *75*, 60–78.

Foxtree, J. E., & Clark, H. H. (1997). Pronouncing the as thee to signal problems in speaking. *Cognition*, *62*(2), 151–167.

Frederiksen, J. R. (1981). Understanding anaphora: Rules used by readers in assigning pronominal referents. *Discourse Processes*, *4*(4), 323–347.

Fridin, M. (2014). Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education. *Computers & Education*, *70*, 53–64.

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological Bulletin*, *133*(4), 694.

Furnas, G., Landauer, T., Gomez, L., & Dumais, S. (1984). Statistical semantics: Analysis of the potential performance of keyword information systems. In *Human factors in computer systems* (pp. 187–242).

Furnas, G., Landauer, T., Gomez, L., & Dumais, S. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, *30*(11), 964–971.

Gaspers, J., Panzner, M., Lemme, A., Cimiano, P., Rohlfing, K., & Wrede, S. (2014). A multimodal corpus for the evaluation of computational models for (grounded) language acquisition. In *Proceedings of EACL Workshop on Cognitive Aspects of Computational Language Learning* (pp. 30–37).

Gatt, A., Krahmer, E., van Deemter, K., & van Gompel, R. P. (2014). Models and empirical data for the production of referring expressions. *Language, Cognition and Neuroscience*, *29*(8), 899–911.

Gernsbacher, M. A., & Hargreaves, D. J. (1988). Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, *27*(6), 699–717.

Gernsbacher, M. A., Hargreaves, D. J., & Beeman, M. (1989). Building and accessing clausal representations: The advantage of first mention versus the advantage of clause recency. *Journal of Memory and Language*, *28*(6), 735–755.

Glenberg, A., & Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9(3)*, 558–565.

Glenberg, A. M., & Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex*, *48*(7), 905–922.

Glenberg, A. M., Witt, J. K., & Metcalfe, J. (2013). From the revolution to embodiment 25 years of cognitive psychology. *Perspectives on Psychological Science*, *8*(5), 573–585.

Gold, K., & Scassellati, B. (2006). Grounded pronoun learning and pronoun reversal. In *Proceedings of the 5th International Conference on Development and Learning.*

Goudbeek, M., & Krahmer, E. (2012). Alignment in interactive reference production: Content planning, modifier ordering, and referential overspecification. *Topics in cognitive science*, *4*(2), 269–289.

Green, A., Hüttenrauch, H., Topp, E. A., & Eklundh, K. S. (2006). Developing a contextualized multimodal corpus for human-robot interaction. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC).*

Grice, H. (1975). Logic and conversation. In *Syntax and semantics: Speech acts* (pp. 41–58). New York.

Grice, M., & Baumann, S. (2002). Deutsche Intonation und GToBI. *Linguistische Berichte*, *191*, 267–298.

Gries, S. (2009). What is corpus linguistics? *Language and Linguistics Compass*, *3*(5), 1225–1241.

Gries, S., & Newman, J. (2013). Creating and using corpora. *Research Methods in Linguistics*, 257–287.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, *82*(B1-B14).

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11*(4), 274–279.

Gross, S., & Krenn, B. (2016). The ofai multimodal task description corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)* (pp. 1408–1414).

Gross, S., Krenn, B., & Scheutz, M. (2016). Multi-modal referring expressions in human-human task descriptions and their implications for human-robot interaction. *Interaction Studies*, *(accepted on 19 Jan 2016).*

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, *21*(2), 203–225.

Gundel, J. K., Bassene, M., Gordon, B., Humnick, L., & Khalfaoui, A. (2010). Testing predictions of the givenness hierarchy framework: A crosslinguistic investigation. *Journal of Pragmatics*, *42*(7), 1770–1785.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 274–307.

Gundel, J. K., Hedberg, N., & Zacharski, R. (2012). Underspecification of cognitive status in reference production: Some empirical predictions. *Topics in Cognitive Science*, *4*(2), 249–268.

Gundel, J. K., Hedberg, N., Zacharski, R., Mulkern, A., Custis, T., Swierzbin, B., ... Watters, S. (2006). *Coding protocol for statuses on the givenness hierarchy.* (unpublished manuscript)

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.

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*(4), 596–615.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, *28*(1), 105–115.

Hardie, A., & McEnery, T. (2010). On two traditions in corpus linguistics, and what they have in common. *International Journal of Corpus Linguistics*, *15*(3), 384–394.

Heeman, P. A., & Allen, J. (1994). Tagging speech repairs. In *Proceedings of the Workshop on Human Language Technology* (pp. 187–192).

Heeman, P. A., & Allen, J. F. (1995). *The trains 93 dialogues.* (Tech. Rep.). Computer Science Department, The University of Rochester: DTIC Document. Retrieved from `http://www.cs.rochester.edu/research/speech/trains.html`

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266.

Horton, W. S., & Rapp, D. N. (2003). Out of sight, out of mind: Occlusion and the accessibility of information in narrative comprehension. *Psychonomic Bulletin & Review*, *10*(1), 104–110.

Huang, C.-M., & Mutlu, B. (2014). Learning-based modeling of multimodal behaviors for humanlike robots. In *Proceedings of the 2014 International Conference on Human-Robot Interaction (HRI)* (pp. 57–64).

Hüwel, S., Wrede, B., & Sagerer, G. (2006). Robust speech understanding for multimodal human-robot communication. In *Proceedings of the 15th International Symposium on Robot and Human Interactive Communication* (pp. 45–50).

Järvikivi, J., van Gompel, R. P., Hyönä, J., & Bertram, R. (2005). Ambiguous pronoun resolution contrasting the first-mention and subject-preference accounts. *Psychological Science*, *16*(4), 260–264.

Jayagopi, D. B., Sheiki, S., Klotz, D., Wienke, J., Odobez, J.-M., Wrede, S., ... Gatica-Perez, D. (2013). The Vernissage corpus: A conversational human-robot-interaction dataset. In *Proceedings of the 8th International Conference on Human-Robot Interaction (HRI)* (pp. 149–150).

Karreman, D., Bradford, G. S., van Dijk, B., Lohse, M., & Evers, V. (2013). What happens when a robot favors someone? how a tour guide robot uses gaze behavior to address multiple persons while storytelling about art. In *Proceedings of the 8th International Conference on Human-Robot Interaction (HRI)* (pp. 157–158).

Kaup, B., & Zwaan, R. A. (2003). Effects of negation and situational presence on the accessibility of text information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 439–446.

Kehler, A. (2000). Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of AAAI/IAAI* (pp. 685–690).

Kelleher, J. D., & Kruijff, G.-J. M. (2006). Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (pp. 1041–1048).

Kendon, A. (2004). *Gesture: Visible action as utterance.* Cambridge University Press.

Kiefer, M., & Barsalou, L. W. (2011). Grounding the human conceptual system in perception, action, and introspection. In *Tutorials in Action Science.* Cambridge: MIT Press.

Kipp, M. (2001). Anvil: A generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)* (pp. 1367–1370).

Kipp, M. (2010). Anvil: The video annotation research tool. *Handbook of Corpus Phonology. Oxford University Press, Oxford.*

Kitagawa, C., & Lehrer, A. (1990). Impersonal uses of personal pronouns. *Journal of pragmatics*, *14*(5), 739–759.

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science*, *30*(3), 481–529.

Kopp, S., Bergmann, K., & Kahl, S. (2013). A spreading-activation model of the semantic coordination of speech and gesture. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society (COGSCI 2013).*

Kowadlo, G., Ye, P., & Zukerman, I. (2010). Influence of gestural salience on the interpretation of spoken requests. In *Proceedings of INTERSPEECH* (pp. 2034–2037).

Krahmer, E., & Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In *Information sharing: Reference and presupposition in language generation and interpretation.* Stanford.

Kranstedt, A., Lucking, A., Pfeiffer, T., Rieser, H., & Wachsmuth, I. (2006). Deictic object reference in task-oriented dialogue. *Trends in Linguistic Studies and Monographs*, *166*, 155.

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, *4*(3), 343.

Kruijff, G.-J. M., Kelleher, J. D., & Hawes, N. (2006). Information fusion for visual reference resolution in dynamic situated dialogue. In *Perception and Interactive Technologies* (pp. 117–128). Springer.

Kruijff, G.-J. M., Lison, P., Benjamin, T., Jacobsson, H., Zender, H., Kruijff-Korbayová, I., & Hawes, N. (2010). Situated dialogue processing for human-robot interaction. In *Cognitive Systems* (pp. 311–364). Springer.

Kügler, F., Smolibocki, B., Arnold, D., Baumann, S., Braun, B., Grice, M., ... Peters, J. (2015). Dima: Annotation guidelines for german intonation. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015).*

Lemaignan, S., Ros, R., Sisbot, E. A., Alami, R., & Beetz, M. (2012). Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction. *International Journal of Social Robotics*, *4*(2), 181–199.

Lindquist, H. (2009). *Corpus linguistics and the description of english.* Edinburgh University Press.

Lozano, S. C., & Tversky, B. (2006). Communicative gestures facilitate problem solving for both communicators and recipients. *Journal of Memory and Language*, *55*(1), 47–63.

MacWhinney, B. (2000). The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database. *Computational Linguistics*, *26*(4), 657–657.

Martell, C., Osborn, C., Friedman, J., & Howard, P. (2002). The FORM gesture annotation system. In *Proceedings of the Multimodal Resources and Multimodal Systems Evaluation Workshop.*

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: structure and processes. *Current Opinion in Neurobiology*, *11*, 194–201.

Mavridis, N. (2007). *Grounded situation models for situated conversational assistants.* Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Mavridis, N. (2015). A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, *63*, 22–35.

Mavridis, N., & Roy, D. (2006). Grounded situation models for robots: Where words and percepts meet. In *Proceedings of the International Conference on Intelligent Robots and Systems* (pp. 4690–4697).

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* Taylor & Francis.

McGuire, P., Fritsch, J., Steil, J. J., Rothling, F., Fink, G. A., Wachsmuth, S., ... Ritter, H. (2002). Multi-modal human-machine communication for instructing robot grasping tasks. In *Proceedings of the International Conference on Intelligent Robots and Systems* (Vol. 2, pp. 1082–1088).

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* University of Chicago Press.

McNeill, D. (2005). *Gesture and thought.* University of Chicago Press.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, *66*(2), B25–B33.

Nilsson Björkenstam, K., & Wirén, M. (2013). Multimodal annotation of parent-child interaction in a free-play setting. In *Proceedings of the 13th International Conference on Intelligent Virtual Agents (IVA).*

Oshima-Takane, Y., Goodz, E., & Derevensky, J. L. (1996). Birth order effects on early language development: do secondborn children learn from overheard speech? *Child Development*, *67*(2), 621–634.

Oshima-Takane, Y., Takne, Y., & Shultz, T. R. (1999). The learning of first and second person pronouns in english: network models and analysis. *Journal of Child Language*, *26*(03), 545–575.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110.

Peirce, C. S., & Buchler, J. (1955). *Philosophical writings of Peirce* (J. Buchler, Ed.). Dover Publications.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(04), 329–347.

Pitsch, K., Lohan, K. S., Rohlfing, K., Saunders, J., Nehaniv, C. L., & Wrede, B. (2012). Better be reactive at the beginning. implications of the first seconds of an encounter for the tutoring style in human-robot-interaction. In *Proceedings of the 21st International Symposium on Robot and Human Interactive Communication* (pp. 974–981).

Prasov, Z., & Chai, J. Y. (2008). What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th International Conference on Intelligent User Interfaces* (pp. 20–29).

Prince, E. F. (1992). The ZPG letter: Subjects, definiteness, and information-status. *Discourse description: diverse analyses of a fund raising text*, 295–325.

Rehm, M., & André, E. (2008). From annotated multimodal corpora to simulated human-like behaviors. In *Modeling communication with robots and virtual humans* (pp. 1–17). Springer.

Reiter, E., Dale, R., & Feng, Z. (2000). *Building natural language generation systems* (Vol. 33). MIT Press.

Rickheit, G., & Wachsmuth, I. (2006). *Situated communication* (Vol. 166). Walter de Gruyter.

Rinck, M., & Bower, G. H. (2000). Temporal and spatial distance in situation models. *Memory & Cognition*, *28*, 1310–1320.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*(2), 217–236.

Roy, D., Hsiao, K.-Y., & Mavridis, N. (2004). Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *34*(3), 1374–1383.

Scheutz, M. (2006). ADE: Steps toward a distributed development and runtime environment for complex robotic agent architectures. *Applied Artificial Intelligence*, *20*(2-4), 275–304.

Scheutz, M., Briggs, G., Cantrell, R., Krause, E., Williams, T., & Veale, R. (2013). Novel mechanisms for natural human-robot interactions in the DIARC architecture. In *Proceedings of AAAI Workshop on Intelligent Robotic Systems.*

Scheutz, M., Schermerhorn, P., Kramer, J., & Anderson, D. (2007, May). First steps toward natural human-like HRI. *Autonomous Robots*, *22*(4), 411–423.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop.*

Schmidt, T., & Schütte, W. (2010). Folker: An annotation tool for efficient transcription of natural, multi-party interaction. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC).*

Schmidt, T., & Wörner, K. (2009). EXMARaLDA - creating, analyzing and sharing spoken language corpora for pragmatics research. *Pragmatics-Quarterly Publication of the International Pragmatics Association*, *19*(4), 565.

Schreitter, S., & Krenn, B. (2013a). Corpus annotation employing a cognitive framework of incremental language understanding. In *Proceedings of the 9th Workshop on multimodal corpora collocated with IVA 2013.*

Schreitter, S., & Krenn, B. (2013b). Phenomena in conveying information during oral task descriptions. In *Proceedings of the Workshop on Embodied Communication of Goals and Intentions Collocated with ICSR 2013.*

Schreitter, S., & Krenn, B. (2014). Exploring inter-and intra-speaker variability in multi-modal task descriptions. In *Proceedings of the 23rd International Symposium on Robot and Human Interactive Communication* (pp. 43–48).

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge university press.

Silverman, K. E., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., ... Hirschberg, J. (1992). TOBI: a standard for labeling English prosody. In *Proceedings of the International Conference on Spoken Language Processing* (Vol. 2, pp. 867–870).

Simmons, W. K., Martin, A., & Barsalou, L. W. (2005). Pictures of appetizing foods activate gustatory cortices for taste and reward. *Cerebral Cortex*, *15*, 1602–1608.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, *12*(2), 153–156.

Staudte, M., & Crocker, M. W. (2009). Producing and resolving multi-modal referring expressions in human-robot interaction. In *Proceedings of the Pre-CogSci Workshop on Production of Referring Expressions.*

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tellex, S., & Roy, D. (2006). Spatial routines for a simulated speech-controlled vehicle. In *Proceedings of the 1st Conference on Human-Robot Interaction (HRI)* (pp. 156–163).

Tenbrink, T., Andonova, E., & Coventry, K. (2008). Negotiating spatial relationships in dialogue: The role of the addressee. *Semantics and Pragmatics of Dialogue (LONDIAL)*, 193.

Tenbrink, T., Eberhard, K., Shi, H., Kuebler, S., & Scheutz, M. (2013). Annotation of negotiation processes in joint-action dialogues. *Dialogue & Discourse*, *4*(2), 185–214.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*, *10*(1), 1–13.

Tomasello, M., & Akhtar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development*, *10*(2), 201–224.

Van Deemter, K., Gatt, A., van Gompel, R. P., & Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, *4*(2), 166–183.

van der Meulen, F. F., Meyer, A. S., & Levelt, W. J. (2001). Eye movements during the production of nouns and pronouns. *Memory & Cognition*, *29*(3), 512–521.

Van der Sluis, I., & Krahmer, E. (2007). Generating multimodal references. *Discourse Processes*, *44*(3), 145–174.

Vollmer, A.-L., Lohan, K. S., Fischer, K., Nagai, Y., Pitsch, K., Fritsch, J., ... Wrede, B. (2009). People modify their tutoring behavior in robot-directed interaction for action learning. In *Proceedings of the 8th International Conference on Development and Learning (ICDL)* (pp. 1–6).

Williams, T., Acharya, S., Schreitter, S., & Scheutz, M. (2016). Situated open world reference resolution for human-robot dialogue. In *Proceedings of the 11th International Conference on Human-Robot Interaction (HRI)* (pp. 311–318).

Williams, T., & Scheutz, M. (2015). Power: A domain-independent algorithm for probabilistic, open-world entity resolution. In *International Conference on Intelligent Robots and Systems (IROS)* (pp. 1230–1235).

Williams, T., Schreitter, S., Acharya, S., & Scheutz, M. (2015). Towards situated open world reference resolution. In *Proceedings of the 2015 AAAI Fall Symposium on AI and HRI.*

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC).*

Yamazaki, A., Yamazaki, K., Ohyama, T., Kobayashi, Y., & Kuno, Y. (2012). A techno-sociological solution for designing a museum guide robot: regarding choosing an appropriate visitor. In *Proceedings of the 7th International Conference on Human-Robot Interaction (HRI)* (pp. 309–316).

Yu, C., Smith, L. B., & Pereira, A. F. (2008). Grounding word learning in multimodal sensorimotor interaction. In *Proceedings of the 30th Conference of the Cognitive Science Society (CogSci)* (pp. 1017–1022).

Zwaan, R. A. (2004). The immersed experiencer: toward an embodied theory of language comprehension. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 35–62). New York: Academic Press.

Zwaan, R. A. (2014). Embodiment and language comprehension: reframing the discussion. *Trends in cognitive sciences*, *18*(5), 229–234.

Zwaan, R. A., & Madden, C. (2005). Embodied sentence comprehension. In D. Pecher & R. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thought* (pp. 224–245). New York: Cambridge Univ. Press.

Zwaan, R. A., & Taylor, L. J. (2006). Seeing, acting, understanding: motor resonance in language comprehension. *Journal of Experimental Psychology: General*, *135*(1), 1.