Multi-modal referring expressions in human-human task descriptions and their implications for human-robot interaction

Stephanie Gross,[1] Brigitte Krenn[1] and Matthias Scheutz[2]
[1]Austrian Research Institute for Artificial Intelligence (OFAI) / [2]Tufts University

Human instructors often refer to objects and actions involved in a task description using both linguistic and non-linguistic means of communication. Hence, for robots to engage in natural human-robot interactions, we need to better understand the various relevant aspects of human multi-modal task descriptions. We analyse reference resolution to objects in a data collection comprising two object manipulation tasks (22 teacher student interactions in Task 1 and 16 in Task 2) and find that 78.76% of all referring expressions to the objects relevant in Task 1 are verbally underspecified and 88.64% of all referring expressions are verbally underspecified in Task 2. The data strongly suggests that a language processing module for robots must be genuinely multi-modal, allowing for seamless integration of information transmitted in the verbal and the visual channel, whereby tracking the speaker's eye gaze and gestures as well as object recognition are necessary preconditions.

Keywords: multi-modal communication, human-robot interaction, reference resolution

1. Introduction

Current interactive, language processing systems are mainly focusing on language. In solely verbal (e.g., remote) situations this approach is reasonable. However, in task descriptions within a shared environment, the role of language is different and other aspects of communication may be just as much as or even more fundamental than language to resolve referring expressions. Therefore, it is critical for future language processing systems in shared environments to take multi-modal references into account.

Interaction Studies 17:2 (2016), **180–210**. DOI 10.1075/is.17.2.02gro
ISSN 1572–0373 / E-ISSN 1572–0381 © John Benjamins Publishing Company

Human teachers use multi-modal communication, most prominently speech, 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; Clark & Krych, 2004; Hanna & Brennan, 2007). While language can be used as the major, possibly even only information channel (e.g., when the communicated information is abstract or when interlocutors communicate remotely), it will often be underspecified and is heterogeneously used by individual speakers (see Furnas et al., 1984, 1987; Brennan, 1996). Language takes a subordinate, guiding role when task-relevant objects and conditions do not have to be inferred from natural language expressions, but can be directly observed. In that case, gestures and gaze of the speaker are often employed as indicative acts during communication (Clark, 2003; Brennan, 2000), even though humans could communicate the intended information through language alone.

For robots to be able to resolve multi-modal referring expressions uttered by human instructors, we thus need to better understand how humans refer to objects in multi-modal task-based communications in order to distil the critical interaction principles that have to be integrated into robotic control architectures.

We hypothesize that in situated referential interaction, human instructors vary vastly in how they structure and present a task to a learner. Our work presented in this paper differs from previous investigations in that we analyse multi-modal referring expressions in situated human-human interactions. We investigate in detail (i) the variation of employed nouns when referring to one specific object, (ii) how often verbal referring expressions are underspecified or omitted, and (iii) the channel(s) which are used to transmit the crucial information for reference resolution. Detailed research questions are listed in Section 3.5.

We then use the analyses of these questions to develop design principles for robotic systems used in natural language interactions with humans.

In the following, we start by reviewing related work and investigate limitations of current reference resolution algorithms to set the stage. Subsequently, we introduce our experimental setup together with the research questions, followed by the presentation and systematic analysis of different communication channels used by human instructors. We specifically aim to isolate essential modality-dependent aspects of interactions that are critical for resolving references to objects in shared scenes. The subsequent discussion section relates the findings to our research questions and points to implications for human-robot interaction. The final conclusion provides a summary of the findings and an outlook for future work.

2. Background and related work

In the work described in this paper, we focus on experimental data from human experiments to inform the development of mechanisms for interpreting referring expressions in robotic architectures. Converging psycholinguistic evidence suggests that perceptions and perspectives of situated, embodied interlocutors are involved during language understanding in humans (Clark & Krych, 2004; Clark, 2003; Brennan, 2000). Computational models aiming at understanding human language thus need to account for its multi-modal complexity. Despite the wealth of empirical research on referring expressions, for instance, in psycholinguistics and the work on developing models in computational linguistics, there has been little mutual influence (see Gatt et al., 2014; Van Deemter et al., 2012, for an overview).

In the following, we will first review literature on multi-modal reference resolution in human-human interaction and then discuss current computational approaches.

2.1 Multi-modal reference resolution in human-human interaction

In natural communication, verbal and non-verbal communication channels together contain the information necessary for understanding the speaker's intention, combining information with various forms of salience. During situated task descriptions, humans communicate by exhibiting, pointing at, placing, and orienting objects, as well as through eye gaze, head nods, and head shakes, all timed with precision (Clark & Krych, 2004).

The ability to refer and identify entities in a shared environment is central to human communication. These referring expressions in natural language can either take the form of a full name (e.g., *Bill*), a pronoun (e.g., *it*) or a description (e.g., *the handle with the green and yellow marker*)[1] (Gatt et al., 2014; Reiter et al., 2000).

The chosen form is influenced 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). Full names are typically used for initial references, while reduced forms, such as pronouns, are used when the object already has high salience (Reiter et al., 2000).

For pronoun resolution, Arnold et al. (2000) found evidence that accessibility information (e.g., proximity, gender) influences referent identification during the initial process of pronoun resolution.

[1] The example is taken from the data collection.

We, however, provide evidence that in situated task descriptions, objects with high visual salience can already be referred to by reduced forms instead of the otherwise typical initial full names even though the uttered information is often insufficient to interpret referring expressions, (e.g., pronouns without antecedents, or pronouns not matching the gender of the antecedent in German).

2.1.1 *Variation in language*

With respect to the average size of the human mental lexicon, the potential for variability in word choices is enormous. Furnas et al. (1987, 1984) called this phenomenon "the vocabulary problem". Brennan & Clark (1996) investigated lexical variability in human-human dialogue. The likelihood that an instructor from one trial chose the same term for the same common objects (e.g., shoes, dogs, cars, fishes) as another instructor in another trial was only 10%. However, within a conversation variability was relatively low. When two people repeatedly discussed the same object, lexical entrainment occurred – they came to use the same terms. For human-robot interaction, the high variability in word choices makes high demands on reference resolution. Extending the study by Brennan & Clark (1996), our analyses will also address lexical variability between and within task descriptions. In contrast to previous studies, we analysed data where mainly one person is speaking while the other one is mainly listening.

In the current paper, we systematically investigate extent and manner of variation of referring expressions between and within situated task descriptions, as well as the amount of underspecified verbal referring expressions.

In order to be able to extract the entities referred to, information transmitted via the visual modality needs to be interlinked with information transmitted via the linguistic modality. In situated communication, visual inferences such as poising, exhibiting, deictic gestures, and eye gaze play an important role in referring to objects and actions (Clark & Krych, 2004). Both gaze and gestures are important cues for establishing joint attention (Prischen et al., 2007; Clark & Krych, 2004; Tomasello & Akhtar, 1995).

2.1.2 *Gesture, gaze and language*

Gestures are an integral part of language, synchronous and co-expressive with speech, and they can be deictic (pointing gestures), iconic, emblems, and beats (McNeill, 2008; Kendon, 2004).

Another potential cue for referential disambiguation is eye gaze (Hanna & Brennan, 2007; Knoeferle & Crocker, 2006; Clark & Krych, 2004). Eye movements are naturally occurring in parallel or even in anticipation of reference production, they are informative, and they are used by addressees as visual cues during reference resolution (Hanna & Brennan, 2007).

In situated task descriptions, participants use linguistic and visual modalities in parallel. Streeck (1993) argues that speech and gesture are linked together by the modality gaze. The important question is how people divide their efforts and information between vocal and visual actions. Clark & Krych (2004) conducted a study in which two participants (instructor and builder) had to collaboratively solve a task. In one scenario, the workspace 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 instructors by exhibiting, poising, pointing at, placing, and orienting blocks, and by eye gaze, head nods, and head shakes. While the two partners were much slower in the second scenario where the workspace was not visible to the instructor, they made more errors in the third scenario. The results provide evidence for the claim that it is essential to consider multi-modal information channels in shared environments.

Depending on the task, speaking, however, might impede understanding. Clark & Krych (2004) argue that participants use vocal and gestural 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 where participants had to search alone, or in pairs with shared gaze, voice, or shared gaze and voice. First, collaborating pairs performed better than solitary searchers, and second, participants were able to solve the task faster in the shared gaze search than in the shared gaze and voice search. These results indicate that verbal descriptions may in some cases also impede understanding. Thus, they might negatively impact task-oriented information exchange, and it is crucial for the design of an artificial agent to include capabilities for detecting and integrating the instructors' eye gaze.

Based on the above evidence, we will take a comprehensive approach in the present study and investigate the interplay of language, gaze and gestures for information transmission.

2.2 Computational approaches to multi-modal reference resolution

Only in the last few years have attempts been made to overcome the gap between experimental research on referring expressions and computational work on algorithms that identify and generate referring expressions (Gatt et al., 2014; Van Deemter et al., 2012). A well-known computational approach to modelling anaphoric reference is *Centering Theory* (Grosz et al., 1995), where 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”. 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 (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, which often have an attention guiding role and not semantic reasons (e.g., Goudbeek & Krahmer, 2012; Arts et al., 2011).

Krahmer & Theune (2002) developed 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, instead of the preceding utterance.

However, these approaches assume that all pronouns can be resolved via antecedents and the 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.

The Givenness Hierarchy by Gundel (2010) consists of six levels of cognitive accessibility: *in focus, activated, familiar, uniquely identifiable, referential,* and *type identifiable.* Each level is contained by all lower levels, so information that is in focus is also activated, familiar, etc. Each level corresponds with a different cognitive status and is realised by a different set of referential expressions, e.g. 'it' is uttered if the speaker believes that the referenced entity is in the interlocutor's focus of attention. Gundel et al. (2006) have also proposed a 'coding protocol', 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 multimodal 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 Grice (1975). By combining these two, they derived the 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. Their 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 multimodal 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.

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.

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, 2009a,b; Fang et al., 2015). However, this is beyond the scope of this paper.

We believe that for language processing in situated contexts, multi-modality is essential. 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.

In this paper, we aim to provide design guidelines to support the development of computational models for human 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. Although previous work has presented bits and pieces of people's verbal and non-verbal referring behaviour in inherently multi-modal situated communication, we are not aware of a study as comprehensive as ours. 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.

3. Data collection experiments and research questions

To investigate multi-modal task descriptions, we designed two data collection experiments. In the first task, an instructor and a learner had to collaboratively move an object. In the second task, an instructor presented and explained to a learner how to connect two parts of a tube and mount it in a box with holdings. Both tasks are short and simple, and can thus be solved based on simple instructions, without requiring additional common sense knowledge for performing the instructions. From the experimental interaction data, we created an annotated data collection that allowed us to investigate human multi-modal interaction patterns.

3.1 Task 1

An instructor and a learner start standing at a table opposite of each other with their goal to collaboratively move an object. On the table between the two participants, there is a board with two handles, see Figure 1. 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, 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.

There is one object with high salience (the board), textures which are part of the main object (the marker or the handle), and a background object (the table). Even though the task focuses on the manipulation of a single object, different referring expressions to the board, the handle or the markers (e.g., "the thing") need to be resolved. In tasks including more objects, multi-modal reference resolution is needed when references are ambiguous. Still, even if references are not ambiguous but underspecified, they need to be resolved by the learner.

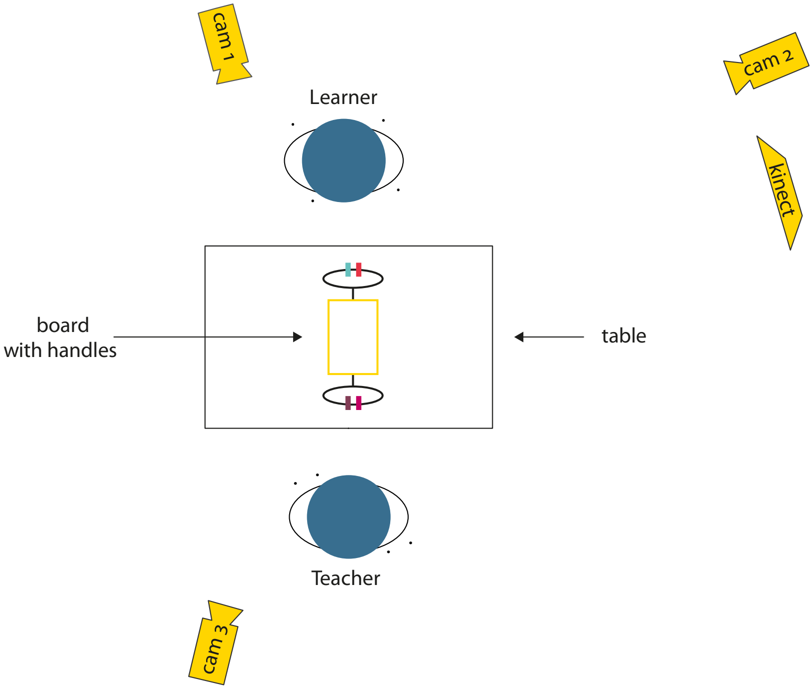

Figure 1. Task 1: Collaboratively moving an object, including the position of the cameras and the Kinect for motion tracking.

3.2 Task 2

In Task 2, a teacher explains and shows a learner how to connect two separate parts of a tube and then mount the tube in a box with holdings. The learner is standing in front of the table at the left side of the teacher (see Figure 2) and is observing 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 teacher. Two coloured markers are attached to the loose part of the tube: a green and yellow one and a red and yellow one, respectively. First, the teacher 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.

As the learner is only observing while the teacher is explaining and conducting the task, she/he has less influence on the task description as in Task 1.

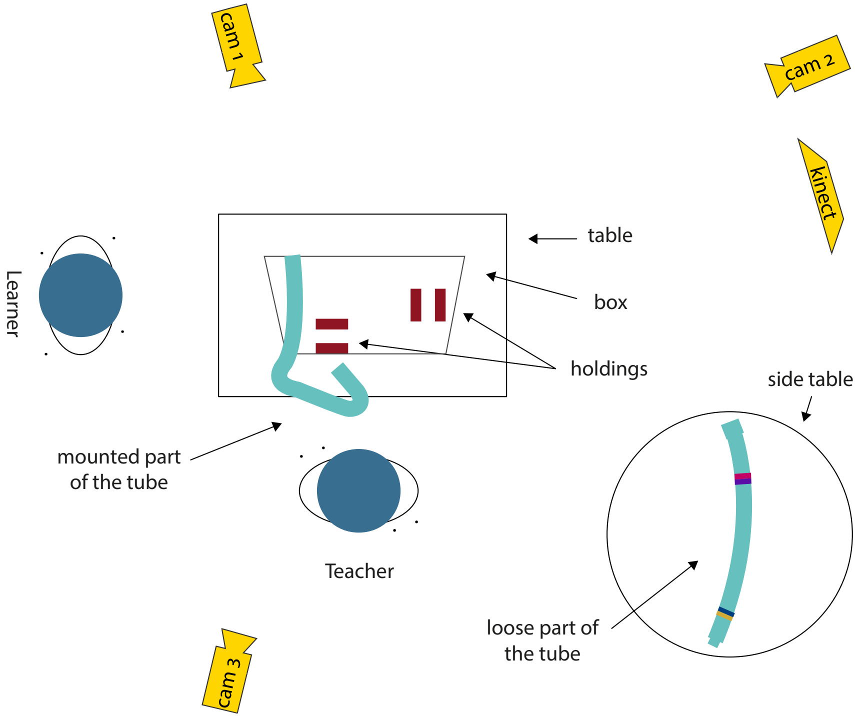

Figure 2. Task 2: Mounting a tube in a box with holdings, including the position of the cameras and the Kinect for motion tracking.

3.3 Data collection

We recorded the utterances of each instructor, a frontal video of the instructor, a frontal video of the learner, and a video of the setting. Moreover, motion was tracked for all activities of the hands, elbows, shoulders, and the head of the instructors. For the recordings, three digital video cameras were used, as well as a wireless microphone worn by the instructor, a receiver, a sound mixer connected to a laptop, and Audacity[2] for audio recording. The teacher's motion was captured via Qualisys Motion Capture Systems[3] and a Kinect sensor, see Figure 1 and Figure 2 for a schematic overview of the setup. The motion tracking data collected via the Qualisys System and the Kinect allow for a more detailed and automatic analysis. Nevertheless, for the study presented in this paper, the audio and video data are sufficient to investigate the multi-modality of referring expressions.

Both task presentations were directed towards a human learner who was told to carefully watch and listen to the explanations of the teacher to be able to pass the information on to a new learner. No restrictions were made in what the learner was or was not allowed to do. However, besides backchannels, only two learners asked questions during the task. 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 instructor to avoid errors that might accumulate in the retellings. The experimenter used the same wording each time, intuitive eye gaze, no gestures. The possibility to use gestures was not mentioned in front of the participants, they were only told to "conduct and explain". Additionally, before each task the instructors received a schematic "cheat sheet" depicting the course of action during the task to reduce their cognitive load.

3.4 Participants and technical tools employed in data analysis

22 subjects 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 for Task 1 (fifteen male and seven female instructors). A subgroup of 16 people of the above mentioned instructors participated in Task 2 and explained the task to a human learner (twelve male and four female instructors). Both groups had an average age of 27 and in both groups one instructor interrupted his/her description in between and started again because he/she forgot how to proceed and started again. Their second run was included in the analyses.

[2] http://audacity.sourceforge.net/

[3] http://www.qualisys.com/

Elan[4] was used for synchronisation purposes of the audio and the video recordings, as well as for annotation. Encoded were (i) the transcription, (ii) the transliteration, (iii) part-of-speech tags employing the TreeTagger (Schmid, 1995), (iv) where the eye gaze of the instructor is directed at, (v) what kind of gesture the instructor employed e.g. for deictic gestures, and where they are directed at, and (vi) the object referred to via language no matter of the lexical choice, e.g. "board" is annotated for "the thing" occurring in the transliteration tier.

3.5 Research questions

The general research question motivating the presented work is to identify mechanisms needed to enable a robot to resolve multi-modal referring expressions occurring in situated task-oriented communication. Therefore, we investigate (i) variation in the choice of nouns denoting one specific object, (ii) underspecification and specification of linguistic referring expressions, and (iii) the role of eye gaze and gestures when uttering referring expressions. Understanding how often multi-modal cues were used and used successfully for resolving references provides design options for the architecture designer (e.g., systems that might not be able to perform gesture recognition will still be able to resolve most references in a task-based context if it turns out that multi-modal cues were consistent with NL descriptions and only employed to help resolve ambiguities). In the following, the three research questions with detailed sub-questions are listed.

RQ1. How large is the variation of expressions referring to individual objects between and within tasks?

a. How high is the inter-speaker variation when referring to individual objects?

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

RQ2. How often are verbal referring expressions underspecified?

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

b. How often are linguistic referring expressions omitted in the utterance?

c. In natural task descriptions, do initial references always contain a description?

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

e. How often are linguistic antecedents of pronouns omitted?

[4] http://tla.mpi.nl/tools/tla-tools/elan/

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

RQ3. What role do different modalities play when referring to objects?

a. How many underspecified verbal referring expressions can be resolved via eye gaze?

b. How many underspecified verbal referring expressions can be resolved via deictic gestures?

c. How often were eye gaze, gestures, and linguistic referring expression contradictory?

d. How many referring expressions could not be identified via language, eye gaze, and gestures?

4. Results

In Task 1, the average task duration was 36 seconds (ranging from 17sec to lmin with 10sec SD), in Task 2 it was 41 seconds (ranging from 18sec to 1min 48sec with 21sec SD). During both tasks, the learners 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 teachers structured and presented the task. In the following, resolutions of object references are presented and discussed along the lines of verbal and non-verbal aspects of referring expressions.

4.1 RQ1 – Variation of referring expressions per object

In Task 1, the main object to be collaboratively manipulated is the board. The other items referred to in the task are the green and yellow markers.

In Task 2, more objects 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 teacher, and green holdings at the left side of the teacher.

In natural human-human interaction, variation can occur due to either underspecification or synonyms. In computational linguistics, synonyms are not a problem as long as they can be handled via lexical databases such as WordNet[5] for English or Universal WordNet[6] for more than 200 languages. For demonstration purposes, we checked whether the different nouns uttered for one object were

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

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

listed in Universal WordNet. Importantly, no synonyms were found in the database because the different nouns were not synonyms but underspecified NPs.

The verbal part of referring expressions contained specific nouns (e.g., “the board”), underspecified nouns (e.g., “the whole”), personal pronouns (e.g., “it”), and deictic adverbs (e.g., “here”) indicating where to place the object – see Table 1 for an overview.

Not surprisingly, how instructors refer to objects does not only depend on the kind of object, but also on the task and the other types of objects available for reference. In Task 1, the board was salient during the whole description. This is reflected in the amount of pronouns occurring as referring expressions. One of the 22 instructors even used pronouns only, without a lexical antecedent when referring to the board. As the marker is on the handle, 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.

Table 1. Summary of all referring expressions to objects of all 22 instructors in Task 1 and all 16 instructors in Task 2. The two objects involved in Task 1 and the seven objects involved in Task 2 are listed. For each object, the chosen surface forms of all referring expressions by all teachers are listed, 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	Referring expressions				Noun variation		Instructors
		specified NP	underspecified NP	pronouns	space deixis	between tasks	within tasks	referring to obj.
Task 1	Board	15	17	58	2	9	2	22/22
	Green and yellow marker	3	16	0	2	8	2	18/22
Task 2	Tube	1	18	17	0	4	2	14/16
	Loose part of the tube	0	15	11	1	7	2	15/16
	Mounted part of the tube	0	8	1	13	5	1	15/16
	Green and yellow marker	11	12	5	2	6	2	15/16
	Yellow and red marker	6	7	2	2	2	2	12/16
	First green holdings	0	12	0	13	10	2	16/16
	Second green holdings	0	14	0	5	8	2	16/16

In Task 2, the seven objects can only be disambiguated via their attributes, and not only via 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 for each of the two objects in Task 1 by all instructors taken together, comprise up to nine different nouns.

In Task 2, up to ten different nouns were used to refer to an individual object (the first 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 2 for all noun phrases referring to the involved objects in Task 1 and Task 2. Only one uttered noun for each object was specific (e.g., *Brett,* German for "board"). The other nouns were either no synonyms (e.g., "beam" – *Balken* in German) or underspecified (e.g., "the whole' – *das Ganze* in German).

Variation occurred between teachers in the choice of nouns designating one specific object, but also within task descriptions. Up to two different nouns were uttered by one teacher within one task to refer to one object, see Table 1.

4.2 RQ2 – Underspecified verbal referring expressions

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

4.2.1 *Verbal part of referring expressions*

When taking a closer look at referring expressions for the board in Task 1, only 22.82% can be resolved based on linguistic information alone, see Figure 3. 16.30% contain the noun *Brett* and 6.52% are pronouns with proximate, congruent and specific antecedents.

Regarding the seven objects in Task 2, fewer pronouns were uttered because these objects were less salient during the task than the board in Task 1. However, in Task 2 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. All together, pronouns were uttered 36 times to refer to the seven objects in Task 2, see Table 1, but only two of these pronouns had a proximate, congruent and specific antecedent.

Spatial deixis "here" (*hier, da* in German) used to refer to objects was mainly uttered when the object was either mounted to, or part of another object, such as

Table 2. Summary of the wording of all 22 participants in Task 1 and all 16 participants in Task 2. 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**
Task 1	Board	‘board’ (*das/dieses/das* (*ganze*) *Brett*) (17;9), ‘thing’ (*das/dieses Ding*) (3;3), ‘object’ (*das Objekt*) (3;2), ‘the whole’ (*das Ganze*) (2;2), (*das / dieses Teil*) (2;2), ‘item’ (*dieser Gegenstand*) (2;1), ‘beam’ (*dieser 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/dieser Griff*) (5;5), (*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), ø (4)
Task 2	Tube	‘tube’ (*-/der/ein Schlauch*) (13;9), ‘the whole’ (*das Ganze*) (2;2), ‘the tube thing’ (*dieses Schlauchteil*) (1), ‘the appendant parts’ (*die zugehörigen Teile*) (1), ‘the two tubes’ (*die zwei Schläuche*) (1), ‘the connected tube’ (*der verbundene Schlauch*) (1), ø (5)
	Loose part of the tube	‘tube’ (*der/dieser/der eine/der andere Schlauch*) (9;9), (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 part/thing’ (*das Teil*) (1), ‘this side’ (*diese Seite*) (1), ø (4)
	Mounted part of the tube	‘tube’ (*der/dieser Schlauch*) (3;3), ‘pipe’ (*das/dieses Rohr*) (2;2), ‘segment of the tube’ (*dieses Teilstück des Schlauches*) (1), ‘end’ (*das Ende*) (1), ‘second tube’ (*der zweite Schlauch*) (1), ø (8)
	Green and yellow marker	‘green and yellow marker’ (*die/diese grün-gelbe/gelb-grüne Markierung*) (11;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 wit 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), ø (1)
	Yellow and red marker	‘red and yellow marker’ (*die rot-gelbe/gelb-rote Markierung*) (6;6), ‘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), ø (4)

Table 2. (*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), ø (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*), ø (4)

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, see Table 1.

One case occurred where the linguistic referring expression was contradictory: one instructor in Task 2 referred to the green and yellow marker as "the green thing" (*das grüne Teil* in German). This referring expression would be more

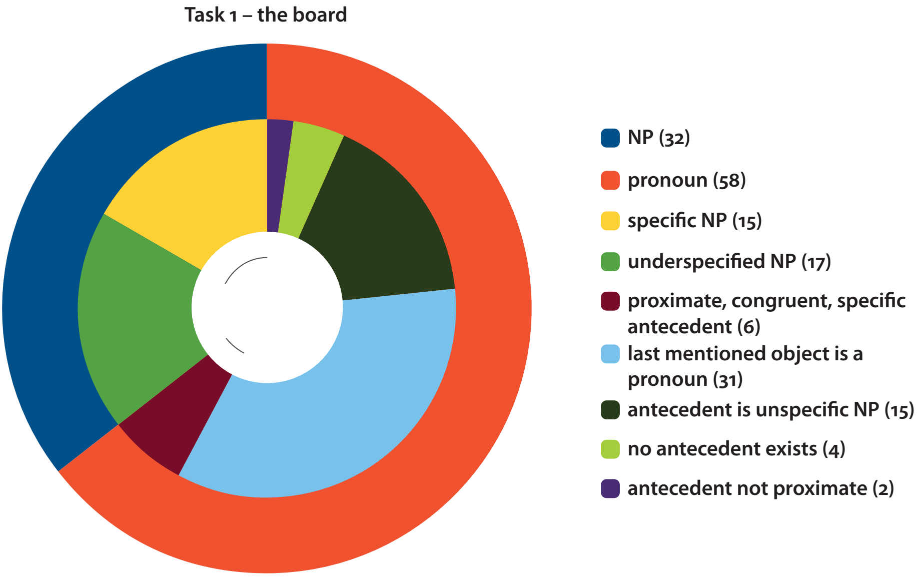

Figure 3. All in all, 32 noun phrases, 58 pronouns and two space deictics were uttered as referring expressions for the tube in Task 2 by 22 instructors. 22.82% can be resolved purely using language (indicated by blue lines in the centre).

appropriate for one of the two green holdings than for the marker. However, the human learner was still able to resolve the reference.

4.2.2 *Verbal part of initial references*

When objects were mentioned for the first time in either tasks, subjects used not only specific noun phrases, but also underspecified noun phrases, space deixis, pronouns, and some instructors even omitted a linguistic referring expression for a specific object altogether.

The first initial reference in Task 1 for the board was "board" by eight (out of 22) instructors. Nine instructors used underspecified noun phrases, four uttered a pronoun and one a space deixis. 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 Unken Hand an den Griff* ***ich mit meiner rechten Hand*** *wir heben es hoch […]* in German). All in all, only 36.36% of the initial

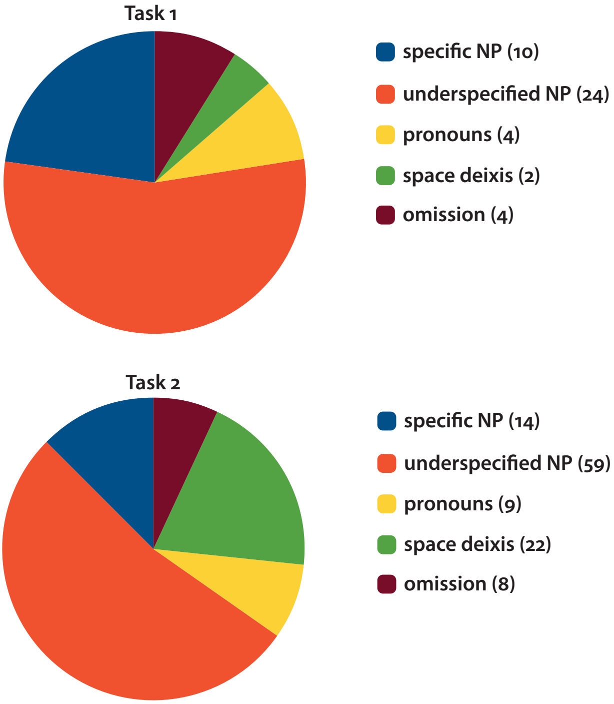

Figure 4. Verbal part of initial references. The pie charts contain all first mentions of the objects (2 in Task 1 and 7 in Task 2). Only initial references by means of specific NPs can be resolved on a linguistic basis only.

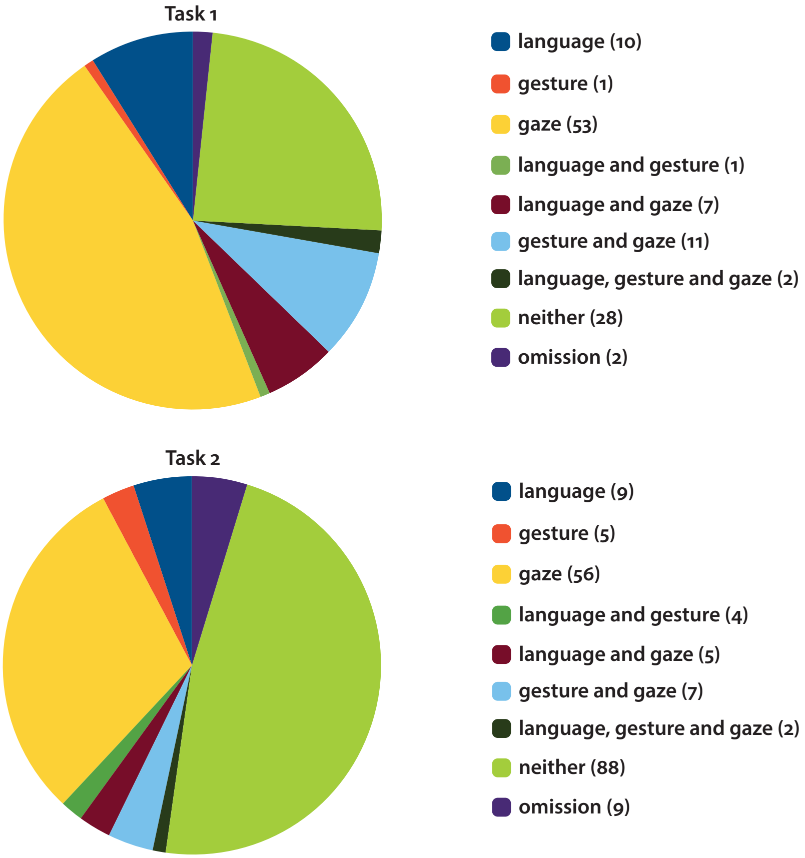

Figure 5. Multi-modal cues of all referring expressions.

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 4). In the majority of the cases, the referents can be resolved by extralinguistic means only.

In Task 2, 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. In both tasks, some instructors omitted referring expressions for objects (see right column of Table 1).

4.2.3 *Pronoun resolution*

Pronouns used were either "it" (*das*),[7] the personal pronoun "it" (*es*), and the demonstrative pronoun "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 1 three instructors used pronouns (3, 1, and 8 pronouns) for referring to the board where the gender of the pronoun was not congruent with the gender of the antecedent. 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). 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. In this context, grammar-based anaphora resolution is deemed to fail and extra-linguistic information is required for reference resolution.

4.3 RQ3 – Multi-modality of referring expressions

Deictic hand gestures were employed by 16 (out of 22) instructors at least once in Task 1 and 6 (out of 16) in Task 2 and frequently used in combination with demonstratives / deictic expressions (e.g., "this end" – *dieses Ende* in German) or space deixis (e.g., "here" – *hier* in German).

In addition to pointing gestures, instructors raised and exhibited an object of attention to the learner. In this respect, Herbert Clark argues that "placing things just in the right manner" (Clark, 2003, p. 243) is an indicative act in which an object is moved into the addressee's attention. In addition to deictic gestures, general communicative gestures (e.g., hands poising above objects in the field of attention), and using fingers for counting and raising the index finger when talking about something important were employed.

Referring expressions were accompanied by deictic gestures in 13.27% of all referring expressions in Task 1 and 10.23% in Task 2. Gestures were the only cue to resolve a reference in 0.88% of all references to objects in Task 1 and 2.84% in Task 2.

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 object referred to in 65.22% of the cases in Task 1. In the other 34.78% of the cases, instead of pointing at the object, they indicated where the

[7] The pronouns *der, die, das* are often referred to as demonstrative pronouns (or d-pronouns), see for instance Ahrenholz (2007), and some linguists argue that in German these pronouns can be interpreted as a second set of personal pronouns, see for instance Lambrecht (1996). In the data collection presented in this paper, the pronouns *das* and *es* were arbitrarily applied, which supports the argumentation by Lambrecht (1996).

salient object should be placed. In Task 2, only one instructor once pointed somewhere else than at the object being verbally referred to.

In the psycholinguistic literature, it is emphasized that the gaze of the speaker is an important cue for disambiguation (e.g., Prasov & Chai, 2008; Hanna & Brennan, 2007; Knoeferle & Crocker, 2006). In Task 1, 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 occurred and only the eye gaze of the instructor was directed at the object. In Task 2, 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 contradictory, but often underspecified. An exception is one instructor who denoted the green and yellow marker with "the green thing" (*das grune Ding* in German). 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, i.e. a specific, proximate, and congruent antecedent.

These results show that eye gaze also independently conveys important information for reference resolution. Different from gestures, the instructor's eye gaze is not applied only for some 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) in Task 1 and 50% in Task 2, see Figure 5.

Based on the collected data, we identified a non-comprehensive list of additional cues necessary to resolve referring expressions:

- *salience* – check whether the last mentioned object is a resolvable reference to the currently mentioned 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)

5. Analysis and challenges

We are aware that the results of human-human interaction can not be transmitted one-to-one to human-robot interaction, as there is also 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; Clark & Krych, 2004; Hanna & Tanenhaus, 2004). We hypothesize that humans expect the most sophisticated communicative behaviour from human interlocutors in contrast to robot interlocutors, no matter of their appearance. By considering human-human interaction, we try to embrace the whole spectrum of human referring behaviour for robotic architecture design, independent of the morphology of the robot.

Due to the detailed analyses based on the research questions, we can extract challenges and provide suggestions for developing a multi-modal reference-resolution mechanism for robots in shared environments with their human interaction partner.

5.1 Challenge 1 – variation of expressions referring to one specific object

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, variability was relatively low and lexical entrainment occurred. In the data presented in this paper, 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.

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

5.2 Challenge 2 – underspecified verbal 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, eye gaze information, and object placement actions 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. According to our data, however, in situated task descriptions visual salience enables the resolution of verbally underspecified references and this issue occurs very frequently, e.g. participants uttered "the thing" for "the board" or "the marker" although there are two differently coloured markers.

5.3 Challenge 3 – multi-modality of referring expressions

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

The fact that one hand of the teachers was occupied during the task description might to some extent explain that relatively few gestures were produced when referring to objects or actions.

5.4 Lessons for agent design

5.4.1 *Variation of expressions referring to one specific object*

A main finding from our situated task interactions is that as language takes more of a scaffolding role, it becomes less reliable than information transmitted via visual channels. In the case of underspecified noun phrases, either more general, less informative nouns were uttered, such as "the thing", "the object", or nouns carrying similar semantic information as the denoted object without being synonyms (e.g., "beam" for "board"). This also raises issues for the learning of new representations. In case the instructor utters "beam" or "thing" instead of "board", the robot should not add "beam" or "thing" as synonyms of "board" to its lexicon. Underspecified and wrong diction should not be learned by the agent. The agent should thus only learn via experience, i.e., extensive exposure to a combination of linguistic and visual input.

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

5.4.2 *Underspecified verbal referring expressions*

Underspecified noun phrases carry too little information and pronouns can not reliably be resolved by means of linguistic information only. Pronoun resolution is a great challenge to human-robot interaction and might hinder the correct interpretation of utterances, especially when disconnected from visual information. In situated task descriptions, antecedents of pronouns are often omitted or underspecified. In German, gender is not a reliable cue for linguistic reference resolution. Neuter is frequently used for pronouns even when the antecedents are male or female.

In general, a large amount of pronouns and underspecified noun phrases need to be resolved via visual channels. For reference resolution, hands and eye gaze should be continuously tracked and dynamically integrated with the utterance processed. In case underspecified noun phrases or pronouns are uttered, the objects at which hands or gaze are directed need to be extracted. Essentially, the detection of underspecified natural language expressions may serve as an attentional mechanism in the robotic architecture that directs the focus of attention on the extra-linguistic visual channels in order to determine the intended referent.

In case the uttered noun phrase is in conflict with the object at which eye gaze and gesture are directed, there are two possibilities for reference resolution: If the noun phrase is specific and identifiable, language is sufficient for reference resolution. If the referring expression is underspecified, the probability is high that the relevant object is the one indicated by eye gaze and/or gesture.

5.4.3 *Multi-modality of referring expressions*

The results show how crucial the detection of eye gaze is during task description. In line with Frischen et al. (2007), Clark & Krych (2004) or Tomasello & Akhtar (1995), the results emphasize that both gaze and gesture are important cues for establishing joint attention and resolving references. Given the closely time-locked referential gaze of the speaker (Griffin, 2001) and the tendency of the listener to follow the speaker's eye gaze (Böckler et al., 2011), monitoring the eye gaze of the speaker might facilitate and, therefore, speed up reference resolution. Thus, eye gaze of the instructor can be used as a predictive cue for what will be uttered. Specifically, by tracking the interlocutor's eye gaze, the visual system of the robotic architecture can attempt to segment and detect objects that might figure in future referential expressions, which will speed up reference resolution and also automatically narrow the set of possible referents (given that interlocutors will almost certainly not refer to objects in the environment they have not gazed at). Moreover, if the robot has what is perceived as "eyes" by a human interactant, then the robot will have to indicate eye gaze following as part of joint attention processes, as it will otherwise risk breaking joint attention patterns (e.g., Chen et al., 2012). For the negotiation of reference is a collaborative process where the speaker proposes a reference and monitors the eye gaze of the listener in order to determine whether it was likely resolved correctly, making incremental adjustments based on that feedback on what to say next or how to expand or correct previous expressions.

Although gestures were rarely the only cue to index objects, taking gestures into account is often necessary for successful reference resolution. However, gestures are also frequently employed to refer to a location while talking about moving an object to a certain location; the same is true of eye gaze. Therefore, the reliability of eye gaze and gestures as cues for object reference resolution also depends on the communicative context and whether the mentioned object is already in focus or not. Hence, the robotic architecture needs to allow for a context-dependent probabilistic integration of multi-modal cues (e.g., as in our recent work on the Givenness Hierarchy (see Williams et al., 2016)).

In addition to language, eye gaze, and gestures of the instructor, additional cues need to be taken into account such as a model of visual and discourse salience, the currently manipulated object, preceding eye gaze, flexible pronoun resolution (e.g., not relying on the gender of the antecedent of objects in German,[8] the reachability of an object, and checking whether the noun or the attribute are sufficient to resolve a reference (e.g., "the green and yellow thing").

6. Conclusion, limitations, and future work

In this paper we presented results from situated task-based interactions between a human instructor and a human learner that highlight the importance of non-verbal communicative cues in situated human task descriptions. Linguistic references were underspecified in the vast majority of cases in our data, and gaze was the most important cue for resolving references in the presented task. As an upshot, linguistic

[8] In contrast, this may not account for agents (see Arnold et al., 2000)

and visual information – especially gaze and gestures – will have need to be incrementally incorporated in a robotic architecture for the robot to be able to resolve referents of underspecified noun phrases or pronouns lacking verbal antecedents.

Some limitations apply to our study: first, the experimental design has the bias of introducing the future instructor by words. If we would not have found any variation, the bias would be a problem. However, the bias is not strong enough, as we found a wide range of variation in word choices. If the future instructor is not introduced by words, we can expect at least the variation we found in our data. This could be addressed in future work.

Secondly, we assume that communication is a bilateral process (Clark & Wilkes-Gibbs, 1986) and instructors' and learners' actions are likely to be coupled. However, our main research interest in this paper is via which channels instructors transmit information, therefore we focus in this paper on the instructors only but plan to include the learners' activities in the future.

Thirdly, the data was collected in German by native speakers. The modalities speech, gaze, and gestures are widely observed behaviours in human interaction across various contexts and cultures (see Benthall et al., 1976; McNeill, 1992). However, it is possible that in a different cultural environment similar experiments might lead to different results. As well as variation according to culture, communicative behaviour can also vary between an experimental setting and the corresponding real world situation the experiment seeks to emulate, or according to the relationship between instructor and instructee. The complexity regarding how humans variously index objects by means of lexical choice, gaze, and gesture is difficult to do justice to. However, the data presented in this paper already shows a wide range of variation, although the experiments were conducted in a laboratory setting and the future instructor was acquainted with the tasks by words. Therefore, the results serve as a good basis for identifying a minimal range of variation.

In future work, we will further investigate the temporal sequence of language, gestures, eye gaze of the instructor and eye gaze of the learner when referring to objects in the shared scene. Additionally, we are interested in prosodic features and their role for structuring information and resolving referring expressions.

Acknowledgements

The first author is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Austrian Research Institute for Artificial Intelligence. The authors would also like to thank the Institute for Information Oriented Control (ITR) at Technical University of Munich and the Cluster of Excellence Cognition for Technical Systems (CoTeSys) for their support with recording the data.

References

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).

Ahrenholz, B. (2007). *Verweise mit Demonstrativa im gesprochenen Deutsch: Grammatik, Zweitspracherwerb und Deutsch als Fremdsprache* (Vol. 17). Walter de Gruyter. doi: 10.1515/9783110894127

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review*, 106(4), 748. doi: 10.1037/0033-295x.l06.4.748

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. doi: 10.1016/S0010-0277(00)00073-1

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, 43(1), 361–374. doi: 10.1016/j.pragma.2010.07.013

Benthall, J., Argyle, M., & Cook, M. (1976). Gaze and mutual gaze. *RAIN*(12), 7. doi: 10.2307/3032267

Böckler, A., Knoblich, G., & Sebanz, N. (2011). Observing shared attention modulates gaze following. *Cognition*, 120(2), 292–298. doi: 10.1016/j.cognition.2011.05.002

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. *Proceedings of International Symposium on Spoken Dialog*, 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). doi: 10.3115/1075218.1075219

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. doi: 10.1016/j.cognition.2007.05.012

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. doi: 10.1037/0278-7393.22.6.1482

Chai, J. Y., Prasov, Z., & Qu, S. (2006). Cognitive principles in robust multimodal interpretation. *Journal of Artificial Intelligence Research (JAIR)*, 27, 55–83.

Chen, Y., Schermerhorn, P., & Scheutz, M. (2012). Adaptive eye gaze patterns in interactions with human and artificial agents. *ACM Transactions on Interactive Intelligent Systems*, 1(2), 13. doi: 10.1145/2070719.2070726

Clark, H. H. (2003). Pointing and placing. *Pointing: Where language, culture, and cognition meet*, 243–268. doi: 10.1075/gest.4.2.08gul

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1), 62–81. doi: 10.1016/j.jml.2003.08.004

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1), 1–39. doi: 10.1016/0010-0277(86)90010-7

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. doi: 10.1016/s0749-596x(02)00001-3

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2), 233–263. doi: 10.1016/0364-0213(95)90018-7

Fang, R., Doering, M., & Chai, J. Y. (2015). Embodied collaborative referring expression generation in situated human-robot interaction. In Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (pp. 271–278). doi: 10.1145/2696454.2696467

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological Bulletin*, 133(4), 691–721. doi: 10.1037/0033-2909.133.4.694

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. 187–212). doi: 10.1002/j.l538-7305.1983.tb03513.x

Furnas, G., Landauer, T., Gomez, L., & Dumais, S. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 50(11), 964–971. doi: 10.1145/32206.32212

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. doi: 10.1080/23273798.2014.933242

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. doi: 10.1111/j.l756-8765.2012.01186.x

Grice, H. (1975). Logic and conversation. In *Syntax and semantics: Speech acts* (pp. 41–58). New York, doi: 10.2307/324613

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, 82(Bl–Bl4). doi: 10.1016/S0010-0277(01)00138-X

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. (2010). Reference and accessibility from a givenness hierarchy perspective. *International Review of Pragmatics*, 2(2), 148–168. doi: 10.1163/187731010X528322

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 271–307. doi: 10.1163/187731010x528322

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. doi: 10.1111/j.l756-8765.2012.01184.x

Gundel, J. K., Hedberg, N., Zacharski, R., Mulkern, A., Custis, T., Swierzbin, B., … Watters, S. (2006). *Coding protocol for statuses on the giveness hierarchy*, (unpublished manuscript)

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. doi: 10.1016/j.jml.2007.01.008

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. doi: 10.1207/sl5516709cog2801_5

Huang, C.-M., & Mutlu, B. (2014). Learning-based modeling of multimodal behaviors for humanlike robots. In Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction (pp. 57–61). doi: 10.1145/2559636.2559668

Huwel, S., Wrede, B., & Sagerer, G. (2006). Robust speech understanding for multimodal human-robot communication. In *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 45–50). doi: 10.1109/roman.2006.314393

Kehler, A. (2000). Cognitive status and form of reference in multimodal human-computer interaction. In Proceedings of the 14th AAAI Conference on Artificial Intelligence (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 (pp. 1041–1048). doi: 10.3115/1220175.1220306

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press, doi: 10.1017/cbo9780511807572

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. doi: 10.1207/sl5516709cog0000_65

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. doi: 10.1515/9783110197747.155

Kruijff, G.-J. M., Lison, P., Benjamin, T., Jacobsson, H., Zender, H., Kruijff-Korbayová, L, & Hawes, N. (2010). Situated dialogue processing for human-robot interaction. In *Cognitive systems* (pp. 311–364). Springer, doi: 10.1007/978-3-642-11694-0_8

Lambrecht, K. (1996). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents* (Vol. 71). Cambridge University Press, doi: 10.2307/417062

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. doi: 10.1007/sl2369-011-0123-x

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. doi: 10.1016/j.jml.2005.09.002

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press, doi: 10.2307/1576015

McNeill, D. (2008). *Gesture and thought*. University of Chicago Press, doi: 10.7208/chicago/9780226514642.001.0001

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27(1), 89–110. doi: 10.1515/ling.l989.27.1.89

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 RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication* (pp. 974–981). doi: 10.1109/roman.2012.6343876

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). doi: 10.1145/1378773.1378777

Reiter, E., Dale, R., & Feng, Z. (2000). *Building natural language generation systems* (Vol. 33). MIT Press, doi: 10.1017/cbo9780511519857

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*.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIG DAT-Workshop*, doi: 10.1007/978-94-017-2390-9_2

Staudte, M., & Crocker, M. W. (2009a). Producing and resolving multi-modal referring expressions in human-robot interaction. In *Proceedings of the Pre-CogSci Workshop on Production of Referring Expressions*, doi: 10.1145/1514095.1514111

Staudte, M., & Crocker, M. W. (2009b). Visual attention in spoken human-robot interaction. In Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction (pp. 77–84). doi: 10.1145/1514095.1514111

Streeck, J. (1993). Gesture as communication i: Its coordination with gaze and speech. *Communications Monographs*, 60(4), 275–299. doi: 10.1080/03637759309376314

Tomasello, M.. & Akhtar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development*, 10(2), 201–224. doi: 10.1016/0885-2014(95)90009-8

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. doi: 10.1111/j.l-8765.2012.01187.x

Van der Sluis, I., & Krahmer, E. (2007). Generating multimodal references. *Discourse Processes*, 44(3), 145–174. doi: 10.1080/01638530701600755

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 (pp. 1–6). doi: 10.1109/devlrn.2009.5175516

Williams, T., Acharya, S., Schreitter, S., & Scheutz, M. (2016). Situated open world reference resolution for human-robot dialogue. In Proceedings of the IEEE/ACM Conference on Human-Robot Interaction (p. forthcoming).

Williams, T., Schreitter, S., Acharya, S., & Scheutz, M. (2015). Towards situated open world reference resolution. In *Proceedings of the 2015 AAAI Fall Symposium on Al and HRI*.

Authors' addresses

Stephanie Gross
Austrian Research Institute for Artificial Intelligence (OFAI)
Freyung 6/6
A-1010 Vienna
Austria

stephanie.gross@ofai.at

Brigitte Krenn
Austrian Research Institute for Artificial Intelligence (OFAI)
Freyung 6/6
A-1010 Vienna
Austria

brigitte.krenn@ofai.at

Matthias Scheutz
Department of Computer Science
Tufts University
200 Boston Avenue
Medford MA 02155
USA

matthias.scheutz@tufts.edu

Biographical notes

Stephanie Gross is a postgraduate researcher at the Austrian Research Institute for Artificial Intelligence (OFAI). She received degrees in cognitive science (M.Sc.) and German philology (M.A.) from the University of Vienna and has worked in various research projects in the field of computational linguistics and human computer interaction at OFAI. Currently she holds a Ph.D. grant from the Austrian Academy of Sciences. Her research interests include multi-modal information transmission in situated human-human and human-robot interaction.

Brigitte Krenn is head of OFAI's Language and Interaction Technologies group. She holds a Ph.D. in computational linguistics from Saarland University Saarbrücken, Germany and a diploma in German language and literature, psychology, philosphy and pedagogy from Karl-Franzens University Graz, Austria. Her research interests lie in combining deep linguistic analysis and shallow corpus-based approaches to text processing and multi-modal communicative interaction, including modelling virtual agent behaviours and multi-modal dialogue.

Matthias Scheutz is a professor of computer and cognitive science in the Department of Computer Science at Tufts University. He received degrees in philosophy (M.A., Ph.D.) and formal logic (M.S.) from the University of Vienna and in computer engineering (M.S.) from the Vienna University of Technology in Austria. He also received a joint Ph.D. in cognitive science and computer science from Indiana University. His current research interests include multi-scale agent-based models of social behaviour and complex cognitive and affective robots with natural language capabilities for natural human-robot interaction.