Exploring Inter- and Intra-speaker Variability in Multi-modal Task Descriptions

Stephanie Schreitter and Brigitte Krenn

Abstract— **In natural human-human task descriptions, the verbal and the non-verbal parts of communication together comprise the information necessary for understanding. When robots are to learn tasks from humans in the future, the detection and integrated interpretation of both of these cues is decisive. In the present paper, we present a qualitative study on essential verbal and non-verbal cues by means of which information is transmitted during explaining and showing a task to a learner. In order to collect a respective data set for further investigation, 16 (human) teachers explained to a human learner how to mount a tube in a box with holdings, and six teachers did this to a robot learner.**

Detailed multi-modal analysis revealed that in both conditions, information was more reliable when transmitted via verbal and gestural references to the visual scene and via eye gaze than via the actual wording. In particular, intra-speaker variability in wording and perspective taking by the teacher potentially hinders understanding of the learner. The results presented in this paper emphasize the importance of investigating the inherently multi-modal nature of how humans structure and transmit information in order to derive respective computational models for robot learners.

I. INTRODUCTION AND MOTIVATION

In human-human task presentation, the interplay of verbal and non-verbal acts brings about the information necessary for the observer or learner to understand. Findings from embodied cognition have shown the importance of multi-modal integration for language comprehension in humans [1], [2]. Thus multi-modal communication must be further investigated for robots to interact with and learn from humans in natural interaction in the future. In the present paper, we investigate which kinds of communicative signals and their variations a robot must be able to perceive and interpret when it is presented with a task. We recorded human-human (HH) and human-robot (HR) teacher-learner dyads to see which information is typically conveyed by the teachers via which channels, focusing on inter- and intra-speaker variability in transmitting information about one and the same task.

A corpus of 22 dyads (16 HH, 6 HR) has been collected to function as an exploratory basis for studying task-oriented communication and the communicative channels involved. In particular, the following research questions were addressed:

- On which channels is relevant information transmitted?
- Which phenomena occur during task descriptions and what is their impact on comprehension?
- What is the inter- and intra-speaker variability in conveying respective information?
- What are the differences in how a task is transmitted between human-human and human-robot dyads?

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. 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 situation assessment, i.e. 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. In the long run, we consider our work as a precondition for future modelling of such components and the resulting corpus as a testbed for respective implementations.

The paper is organized as follows: The related work section is followed by a section where the data collection procedure is described, an analysis and a discussion section investigating the phenomena present in the corpus.

II. RELATED WORK

HH communication is a bilateral process, in which speaking and listening together form a joint activity. Speakers monitor not only their own actions, but also those of the addressees [3]. This is especially relevant in the area of HR interaction. Several studies have been directed to the appearance and morphology of the robot and resulting impact on the HR interaction. Influencing parameters 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, e.g. [4], [5]. In our study, we explore multi-modal variations on the teacher side in transmitting task information to humans and a robot. Identifying these recurring phenomena and incorporating them when developing future robot architectures has high potential for enhancing HR communication.

Psycholinguistic studies guide our investigations of HR interaction. In task descriptions, human learners use information from the visual and the communicative context as cues to disambiguate and understand what is presented by the teacher. Converging psycholinguistic evidence suggests

Stephanie Schreitter and Brigitte Krenn are affiliated with the Austrian Research Institute for Artificial Intelligence, 1010 Vienna, Austria firstname.lastname@ofai.at; Stephanie Schreitter is also affiliated with the University of Vienna, 1010 Vienna, Austria

that during human language understanding perceptions and perspectives of situated, embodied interlocutors are involved. Language processes are coordinated interpersonally but also within the mind of an individual, including the ability to mentalize about another person's mental state (Theory of Mind) and respond rapidly and automatically [6]. Clark and Krych [3] emphasize that HH dialogue is a bilateral, opportunistic, and multi-modal process where common ground is continuously updated. They argue that in dialogue, participants use vocal and gestural modalities in parallel and that the visual modality is faster and more secure than the auditory modality for certain types of communication. This is in accordance with our findings that information transmitted via verbal and gestural references to the visual scene was more reliable than the actual naming of objects and actions.

Gestures are synchronous and co-expressive with speech, cf. [7], [8]. In addition to communicative gestures, the perceived gaze direction of the teacher can be used as an indicator of another person's future actions [9]. This was also clearly found in the present data.

Some attempts have been made to develop mechanisms for robot architectures that support multi-modal understanding, e.g. [10], [11], [12], [13]. The present work contributes to further sharpening the picture of which phenomena necessarily must be dealt with.

III. DATA COLLECTION

Letting several humans do and explain one and the same task helps to better understand how humans structure and present information and what a robot would have to deal with when it were in the learner's position.

A. Setting and Procedure

The objective of the data collection was to create a corpus where a teacher explained and showed to a learner how to mount a tube. The task is borrowed from a robot setting and has not much in common with everyday life. This has the advantage that no everyday knowledge is needed for understanding. Recorded were: the utterances of each teacher, a frontal video of the teacher, a frontal video of the learner, a video of the setting, and motion was tracked. Three digital video cameras were used, as well as a wireless microphone (attached to the teacher), a receiver, a sound mixer connected to a laptop, and Audacity[1] for audio. Motion was captured via Qualisys Motion Capture Systems[2] and a Kinect sensor. See Fig. 1 for an overview of the setup.

Scenario: The task to be described is short and simple. It contains a grasp for the loose part of the tube on the right hand-side of the teacher, performed with the right hand. It must be connected at the green and yellow marker with another part of the tube which is already attached to the box. The tube then must be placed between two green holdings at the green and yellow marker, grasped at another coloured marker and put between another pair of green holdings.

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

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

Human-Human (HH) Dyads: The observer was told to carefully watch and listen to the explanations to be able to pass the information on to a new learner. In the following trial, the learner became the new teacher. After every fourth trial, a calibration trial was introduced where the experimenter functioned as teacher to avoid the Chinese whisperer's effect. Additionally, the teachers received a 'cheat sheet' depicting the course of action during the task.

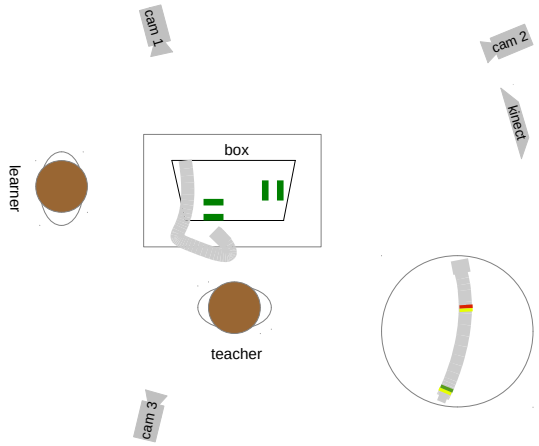

Fig. 1. The task setup for mounting the tube.

Human-Robot (HR) Dyads: In the HR dyad, a robot research prototype developed at the Institute of Automatic Control Engineering at the Technical University in Munich was employed. The robot is of human-size height and equipped with an omni-directional mobile platform, two anthropomorphic arms, and a pan-tilt unit on which Kinect sensors are mounted, see Fig. 4. All robot actions (i.e. head movements and verbal feedback) were controlled by a human wizard. Empirical evidence has shown that head-movement and verbal backchannel feedback by listeners during utterances communicate understanding and are expected by human speakers [14]. Therefore, the 'head' of the robot (the Kinect) was controlled by a human during the task descriptions and directed either towards the setting or towards the face of the teacher. In addition, the MARY Text-to-Speech Synthesis platform[3] was employed to give verbal feedback (for technical reasons, verbal feedback worked for five of the six participants). In the HR setting, the task was explained to each teacher by the experimenter in exactly the same wording and they also received a 'cheat sheet'.

The resulting corpus comprises 22 German recordings. In 16 recordings, the task descriptions are directed towards a human learner, and in six towards a robot learner.

B. Participation

22 people working or studying at the Technical University Munich or the Ludwig-Maximilians-University Munich participated in the data collection activity as teachers. Their mother tongue was German. Twelve male and four female

[3]http://mary.dfki.de/

teachers with an average age of 27.19 explained the task to a human learner. One teacher in the HH setting had to conduct the task twice because she forgot how to mount the tube. Additionally, three male and three female participants with an average age of 25.83 explained the task to a robot.

IV. DATA ANALYSIS

The audio and video recordings were synchronised in Elan[4]. The transcriptions of the utterances were analysed as well as the three videos. The data analysis focuses on identifying (i) prevalent channels on which information relevant for task understanding is transmitted, (ii) recurring inter- and intra-speaker variability during task descriptions, and (iii) information necessary to resolve ambiguities. The phenomena identified will be described in the following.

While motion tracking via Qualisys System and Kinect allows for a more detailed and automatic analysis, the videos are sufficient for investigating which communicative gestures are employed and how often. Therefore we did no include the motion tracking data in this analysis.

In addition to the detailed multi-modal analysis, a questionnaire was employed in the HR setting. It was used to provide an insight into the participants' acquaintance with state-of-the-art in robotics and speech synthesis, as this might influence their assessment of the robot and the interaction in general. The participants were asked

- whether they have worked with robots before and if yes, in which scope.
- 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.

V. RESULTS AND DISCUSSION

Apart from the characteristics of spoken language, the multi-modal qualitative analysis of the data revealed the following: (i) variation in wording regarding objects and actions, as well as omissions of lexical referents, (ii) patterns of use of verbal references and/or communicative gestures for directing the learner's attention, (iii) temporal structuring of the task by verbal means, (iv) variation in the speaker perspective taken by the teachers, and (v) eye gaze as a predictor for upcoming areas for attention.

A. Human-Human Communication

On average, the task duration was 41 seconds (25sec - 1min 48sec; SD: 21sec). Even though the task was quite simple there was considerable variation in how teachers structured and presented the task. In the following, prevalent phenomena are presented and discussed.

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

Fig. 2. Human learner. A teacher is mounting a tube in a box with holdings.

1) Characteristics of spoken language: Several characteristics typical for spoken language are present in the data. Occurring phenomena and examples are:

- insertions (11)[5] – 'äh', 'like this' (*so*)[6]
- sentence fragments (9) – 'this is somehow' (*das ist irgendwie*), 'put the green and yellow marker' (*steckst die grün-gelbe Markierung*)
- repairs (6) – 'red eh blue and yellow marker' (*rot äh blau-gelben Marker*), 'so we I take' (*so wir ich nehm*)
- contractions (6) – 'through the' (*durchs*, 'durch das'), 'explain it' (*erklaers*, 'erklaere es')
- repetitions (2) – 'the the' (*die die*)
- dialect (2) – 'no this is not working' (*na des hebt net* for 'nein das hält nicht') – dialect was used in meta comments when the utterance was not directed to the learner, e.g. when connecting the parts of the tube did not work.
- omission (1) – 'takes one (person) tube' (*nimmt man Schlauch* for 'nimmt man den Schlauch')

In HH and HR interaction, disfluencies can inhibit correct interpretation, especially when disconnected from the visual information.

2) Variations in wording: Relevant entities occurring in the task are: two parts of a tube, two pairs of green holdings, two markers, and the right hand of the teacher. Relevant actions are: grasping the tube, connecting the parts of the tube, and putting the tube between the green holdings at the two markers, see Table 1.

The analysis of variations 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. Additionally, the unspoken needs to be grounded in the scene, e.g. the first green holdings were not mentioned by five teachers, and the verbal expressions for the spatial perspectives has to be visually resolved, e.g. one teacher named the second green 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 teacher has to be able to resolve what is uttered to what is in the world, cf. [15].

[5]The number of teachers uttering the characteristics.

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

TABLE I

SUMMARY OF THE WORDING IN HUMAN-HUMAN AND HUMAN-ROBOT DYADS.

Referent	**Wording in HH dyads**	**Wording in HR dyads**
Tube	'tube' (*Schlauch*) (14), 'the whole' (*das Ganze*) (2), 'pipe' (*Rohr*) (2), 'loose pipe' (*lose Rohr*) (1), '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)	'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*) (8), ∅ (8)	'right hand' (*rechte Hand*) (4), ∅ (2)
Green and yellow marker	'green and yellow marker' (*grün-gelbe Markierung*) (5), (*gelb-grüne Markierung*) (4), '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-gruene Verbindung*) (1), 'green and yellow part' (*grün-gelbe Teil*) (1), 'green and yellow section' (*grün-gelbe Abschnitt*) (1), 'this side' (*diese Seite*) (1), 'green thing' (*grüne Teil*) (1), ∅ (1)	'green and yellow marker' (*grün-gelbe Markierung*) (1), (*gelb-grüne Markierung*) (1), 'yellow and green connection' (*gelb-gruene Verbindung*) (1), 'end with the green and yellow glue' (*Ende mit dem grünen und gelben Kleber*) (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' (*Schlauch*) (4), 'pipe' (*Rohr*) (1), 'segment of the tube' (*Teilstück des Schlauches*) (1), 'second tube' (*zweite Schlauch*) (1), ∅ (9)	'tube' (*Schlauch*) (2), 'tube at the mounting' (*Schlauch bei der Befestigung*) (1), 'mounted tube' (*befestigter Schlauch*) (1), 'pre-assembled tube' (*vorgefertigter Schlauch*) (1), 'other part of the tube' (*andere Teil vom Schlauch*) (1)
First green holdings	'mounting' (*Befestigung*) (1), 'this side' (*diese Seite*) (1), 'holding' (*Halterung*) (1), '(right) first holding' (*erste Halter*) (1), (*erste Halterung*) (1), (*rechte erste Halterung*) (1), 'first barrier' (*erste Hindernis*) (1), 'green thing' (*grüne Ding*) (1), 'two blocks' (*beiden Klötze*) (1), 'right green marker' (*rechte grüne Markierung*) (1), 'right channel' (*rechte Kanal*) (1), 'appliance' (*Vorrichtung*) (1), ∅ (5)	'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 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)	'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*) (5), (*gelb-rote Markierung*) (1), 'the yellow and red (section/part)' (*der gelb-rote (Abschnitt/Teil)*) (2), '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), 'the red one' (*das rote*) (1), ∅ (4)	'red and yellow marker' (*rot-gelbe Markierung*) (1), (*rote und gelbe Markierung*) (1), (*gelb-rote Markierung*) (1), 'second marker, the red and yellow one' (*zweite Markierung das rot-gelbe*) (1), 'other end' (*andere Ende*) (1), 'red and yellow connection' (*rot-gelbe Verbindung*) (1)
Take	'take' (*nehmen*) (11), 'work' (*arbeiten*) (1), 'grasp' (*greifen*) (1), ∅ (3)	'take' (*nehmen*) (6)
Assemble	'assemble' (*hineinstecken*) (10), (*zusammenstecken*) (2), (*anstecken*) (1), (*drinnen stecken*) (1), 'connect' (*verbinden*) (4), 'combine' (*kombinieren*) (1), ∅ (1)	'assemble' (*stecken*) (5), (*hineinstecken*) (1)
Put	'put through' (*durchlegen*) (4), (*durchführen*) (2), 'insert' (*hineinlegen*) (3), (*einführen*) (1), (*einlegen*) (1), (*hinein kommen*) (1), (*hineintun*) (1), (*stecken*) (1), (*zum Liegen kommen*) (2), '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)	'put' (*legen*) (4), 'assemble' (*stecken*) (1), 'put through' (*durchlegen*) (1), (*hindurchlegen*) (1), (*durch tun*) (1)

Variation in wording may in part depend on the prototypicality of an object or action. While the tube is a prototypical object, the green barriers are not, and thus cause more variation in wording. Regarding actions, the possibility of prefixing 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*). Additionally, this action is more relevant for achieving the task than for example grasping the tube and teachers may put more effort into accurate explanations. However, some verbs do not precisely describe the shown action, e.g. 'thread' (*durchfädeln*) and some expressions are more specific than others, e.g. 'insert' (*hineinlegen*)' versus 'that it goes inside' (*dass es hineingeht*) or 'green and yellow marker' (*grün-gelbe Markierung*) versus 'this side' (*diese Seite*). Also intra-speaker variation in wording is a factor, e.g. one teacher used two different words for denoting the tube: 'the tube' (*der Schlauch*) and 'the thing' (*das Teil*); another teacher uttered four different verbs for putting the tube between the green holdings: 'insert' (*hinein kommen*), 'gets' (*kommen*), 'put through' (*durchlegen*), 'put' (*legen*).

3) Time Markers: All teachers used verbal time markers to structure the task, e.g. 'then' (*dann*), 'subsequently' (*anschließend*), 'now' (*jetzt*). Additionally, almost all teachers (14) took a step back when they had finished the task. This finalising action demonstrated that the task was done. One teacher alternatively made a posture shift and one stopped in the middle of the movement because the tube fell apart.

4) Teachers' perspective: The teachers' expression of perspective varied in person not only between but also within speakers, see Fig. 3 for a summary. Two speakers used 1^{st}, three 2^{nd} person singular, three 1^{st} person plural and one the indefinite pronoun in indicative mood. One teacher used 2^{nd} person singular in subjunctive mood. The other six teachers

changed the grammatical person during their explanation. Three of these six participants 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 the indefinite pronoun and 1^{st} person singular, between 1^{st} and 2^{nd} person singular and between 1^{st} person plural, 1^{st} person singular, indefinite pronoun, and passive voice 'the marker comes in the first holdings' (*die Markierung kommt in die erste Halterung*).

In three cases, taking the perspective of 2^{nd} person singular by the teacher was mistaken by the respective learners to become active in conducting the task.

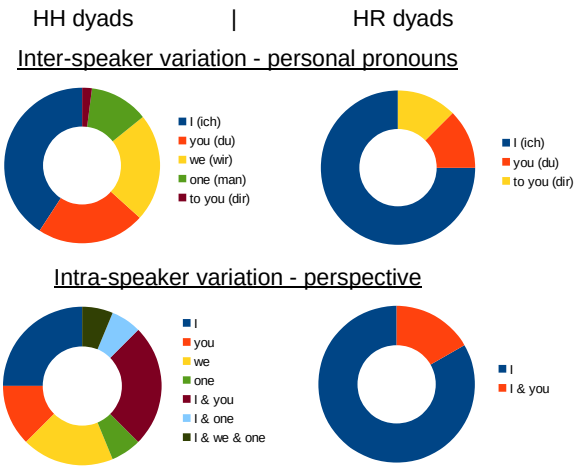

Fig. 3. Top: all personal pronouns used are illustrated (16 HH dyads on the left, 6 HR dyads on the right). Bottom: variation of perspective within speakers is depicted (again HH dyads on the left, HR dyads on the right).

5) Verbal and gestural references to the scene: Deictic pointing gestures were employed by six teachers to guide the attention of the learner. These gestures were frequently used in combination with demonstratives / deictic expressions, e.g. 'this end' (*dieses Ende*), or verbal references to the scene, e.g. 'here' (*hier*). Three teachers raised and exhibited an object of attention to the learner. In this respect, Herbert Clark argues that "placing things just in the right manner" ([16], 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 by one person each.

All teachers uttered verbal references to the visual scene. These references referred to locations, e.g. 'here' (*hier, da*), the manner in which the task was conducted, e.g. 'like this' (*so*), or 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*).

6) Eye gaze: Except for one participant, all teachers looked at their respective learner before, during and/or after the task. Three looked at the observer only once at the beginning, two during the task, and four at the end. Six teachers frequently looked at the learner, up to six times.

Ten teachers looked at the object or location where the attention was going to be directed to before they talked about it. Therefore their eye gaze was predictive for where to the attention would go next. Endsley [17] argues that situation awareness allows for projection of future situations.

B. Human-Robot Communication

The average task duration in the HR dyads was 41 seconds (33sec - 1min 1sec; SD: 11sec).

Fig. 4. Robot learner. A teacher is mounting a tube in a box with holdings.

1) Characteristics of spoken language: The characteristics found were similar to those found 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')

2) Variations in wording: Except for the right hand, each object and action was mentioned by each teacher, see Table 2. However, less variation occurred than in the human-human interactions, e.g. for 'put'. Less intra-speaker variation in wording and employing more prototypical words might be due to the adaption of the human teachers to the robot learner.

3) Time Markers: Again, all teachers used verbal time markers to structure the task in sub-tasks, e.g. 'then' (*dann*), 'now' (*nun*). Only two teachers took a step back when they had finished the task and three conducted a more restrained posture shift. The teacher for whom the verbal feedback from the robot failed did not change her posture at all after finishing the task.

4) Teachers' perspective: The perspective taken by the teacher in the HR interaction was 1^{st} person singular for all teachers. Only one teacher changed to 2^{nd} person singular after a short pause in which she had problems in connecting the two parts of the tube. This predominant use of 1^{st} person singular is rather surprising and needs further investigation.

5) Verbal and gestural references to the scene: Five out of six teachers used deictic gestures, and the sixth person moved his fingers to enable sight on the coloured markers.

These gestures were also used in combination with demonstratives / deictic words, e.g. 'this tube' (*diesen Schlauch*),

or verbal references to the scene, e.g. 'here' (*hier*). Three teachers exhibited the tube or their right hand in order to emphasize that the task has to be conducted with the right hand or that the tube is now in the focus of attention.

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

6) Eye gaze: Each teacher looked at the robot learner at the beginning. All in all, they looked at the robot during the presentation 2 to 11 times. Only the eye gaze of one teacher allowed for inferring the upcoming focus of attention.

C. Questionnaire

All participants were naive to robots. Only one out of six participants had contact to robots before as a participant in a user study. Answering the question whether they thought that an algorithm or a human was controlling the robot's head movements, five out of six participants had the impression the robot's head was controlled by an algorithm. One of them had participated in a study on speech synthesis before. For all others, speech synthesis was new. Overall, the mean value for evaluating the naturalness of the interaction with the robot was 3.33 (1 very natural, 5 not natural at all; SD: 1.21).

VI. CONCLUSION AND FUTURE WORK

The information necessary for reproducing the task was transmitted by the teachers via verbal descriptions together with exhibiting objects, poising, pointing at objects or locations, placing objects, and eye gazes. Aspects complicating comprehension of the task were (i) characteristics of spoken language, (ii) inter- and intra-speaker variations in wording, and (iii) the different perspectives taken by the teachers, and the varying perspectives taken by individual teachers. Signals facilitating understanding were (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 predicting the upcoming area of attention. These signals were more reliable than the lexical expression of entities, actions, and speaker perspective.

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. The question arises where these differences come from. Before explaining the task to the robot, the teachers 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. Teachers also might have thought of the robot more as a camera than as a learner and thus chose 'I' (*ich*) as the preferred perspective. However, the number of eye gazes toward the robot learner was on average higher than toward the human learner, which speaks against the assumption of the robot being perceived as a camera. These differences need to be further investigated to allow for a deeper understanding of the effects. Additionally, future work will focus on information structure and prosody in task descriptions, differences in goal- and action-based explanations, and the assessment of the teachers' motion data.

VII. ACKNOWLEDGMENT

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 the robot and with recording the data.

REFERENCES

[1] M. Kiefer and L. Barsalou, "Grounding the human conceptual system in perception, action, and introspection," in *Tutorials in action science*, MIT Press, 2011.

[2] R. A. Zwaan and L. Taylor, "Seeing, acting, understanding: motor resonance in language comprehension," *Journal of Experimental Psychology*, vol. 135, pp. 1-11, 2006.

[3] H. H. Clark and M. A. Krych, "Speaking while monitoring addressees for understanding," *Journal of Memory and Language*, vol. 50, no. 1, pp. 62-81, 2004.

[4] A-L. Vollmer, et al. "People modify their tutoring behavior in robot-directed interaction for action learning," in *Proceedings of the 8th International Conference on Development and Learning*, 2009.

[5] K. Pitsch, K. S. Lohan, K. Rohlfing, J. Saunders, C. L. Nehaniv, and B. Wrede, "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 22nd IEEE International Symposium on the Robot and Human Interactive Communication*, 2012.

[6] S. E. Brennan, A. Galati, and A. K. Kuhlen, "Two minds, one dialog: Coordinating speaking and understanding," *Psychology of Learning and Motivation*, vol. 53, pp. 301-344, 2010.

[7] D. McNeill, "Gesture and thought," University of Chicago Press, Chicago, 2005.

[8] K. Bergmann, "The production of co-speech iconic gestures: empirical study and computational simulation with virtual agents," PhD Thesis, Bielefeld Univ., Bielefeld, Germany, 2012.

[9] A. Frischen, A. P. Bayliss, and S. P. Tipper, "Gaze cueing of attention: visual attention, social cognition, and individual differences," *Psychological bulletin*, vol. 133, no. 4, pp. 694-724, 2007.

[10] M. Scheutz, G. Briggs, R. Cantrell, E. Krause, T. Williams, and R. Veale, "Novel mechanisms for natural human-robot interactions in the DIARC architecture," in *Proceedings of AAAI workshop on Intelligent Robotic Systems*, 2013.

[11] S. Kopp, K. Bergmann, and S. Kahl, "A spreading-activation model of the semantic coordination of speech and gesture," in *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, 2013.

[12] S. Hüwel, B. Wrede, and G. Sagerer, "Robust speech understanding for multi-modal human-robot communication," in *Proceeding of the 15th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 948-954, 2006.

[13] S. Lemaignan, R, Ros, E. Sisbot, R. Alami, and M. Beetz, "Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction," *International Journal of Social Robotics* vol. 4, pp. 181-199, 2012.

[14] K. Eberhard, M. Spivey-Knowlton, J. Sedivy, and M. Tanenhaus, "Eye movements as a window into real-time spoken language comprehension in natural contexts," *Journal of Psycholinguistic Research*, vol. 24, pp. 409-436, 1995.

[15] R. Cantrell, M. Scheutz, P. Schermerhorn, and X. Wu, "Robust spoken instruction understanding for HRI," *Proceedings of the 2010 Human Robot Interaction Conference*, 2010.

[16] H. H. Clark, "Pointing and placing," in *Pointing. Where language, culture, and cognition meet*, S. Kita, Eds., Hillsdale, NJ, Erlbaum, pp. 243-268, 2003.

[17] M. R. Endsley, "Theoretical underpinnings of situation awareness: A critical review," in *Situation awareness analysis and measurement*, M. R. Endsley and D. J. Garland, Eds., Erlbaum, 2000.