The OFAI Multimodal Task Description Corpus

Stephanie Gross, Brigitte Krenn
Austrian Research Institute for Artificial Intelligence
Freyung 6, 1010 Vienna, Austria
{stephanie.gross, brigitte.krenn}@ofai.at

Abstract

The OFAI Multimodal Task Description Corpus (OFAI-MMTD Corpus) is a collection of dyadic teacher-learner (human-human and human-robot) interactions. The corpus is multimodal and tracks the communication signals exchanged between interlocutors in task-oriented scenarios including speech, gaze and gestures. The focus of interest lies on the communicative signals conveyed by the teacher and which objects are salient at which time. Data are collected from four different task description setups which involve spatial utterances, navigation instructions and more complex descriptions of joint tasks.

Keywords: multimodal task description, open world reference resolution, multimodal human-robot interaction

1. Introduction

Future robots are expected to be present in people's homes, collaborate with and support their human users in various everyday activities and tasks including mobility, manipulation, personal care, fetch and carry support in the household, and so forth. Bringing robots into real-world and thus open environments requires, amongst many other aspects of research and technological development, the creation of artificial learners that can learn from being exposed to task descriptions given by a human tutor. These task descriptions are multimodal inputs where the artificial learner, the robot, is exposed to the full bandwidth of modalities forming natural human-human communication in shared environments. This includes natural language utterances with references that are linguistically underspecified, the omissions of referents in the utterance at all, the use of pronouns without antecedents, and so forth, however, combined with body gestures such as hand-arm gestures, head movements, eye gaze, posture, etc. Taken together, the multimodal input stream, the objects and actions in a given task convey the information a listener needs in order to fully understand the verbal task description. The task scenarios for the OFAI-MMTD corpus were designed with the goal of collecting data based on which the following can be studied:

- How speakers refer to objects, navigation paths etc. in a particular task setting.
- What the interplay of gesture, eye gaze, and language is in a particular task demonstrated by a human tutor.
- How large the inter/intra-speaker variation is when referring to objects.
- How often linguistic expressions are omitted such as verbs, pronouns, nouns.
- What the (multimodal) cues are to prime a listener/learner to pay attention to objects, paths and actions.

Several corpora comprising instructor-learner interactions have been collected, the majority of which are caregiver-child interactions. A large resource is the CHILDES data base which serves as a central repository for first language acquisition data. Moreover, Björkenstam and Wirén (2013) , as well as Yu et al. (2008) collected and annotated multimodal caregiver-child interactions. By contrast, we are interested in the variation of the communication signals between, but also within task descriptions. Thus, we need different people explain the same task, in order to better understand how humans naturally structure and present information.

In line with the corpus developed by Gaspers et al. (2014) we attempt to investigate the multimodal interaction with respect to the communicated tasks. The corpus developed by Gaspers et al. was designed to support the evaluation of computational models addressing several language acquisition tasks, in particular the acquisition of grounded syntactic patterns. Thus, they predefined objects and actions, reappearing several times. In our corpus, the focus is on the task in general – not on specific actions including specific objects – to capture differences in how people structure and present information.

The task scenarios for the data collection and the technical setups for data recording are presented in Section 2. The annotation tiers for the MMTD corpus V1 gold standard are described in Section 3. The paper concludes with examples for an early use of the annotated corpus and an outlook to future work (Section 4.).

2. OFAI-MMTD Corpus – Data Collection

Data were collected from four different task scenarios where in individual teacher-learner pairings a teacher explains and shows four different tasks to a learner. The idea behind letting different people explain the same tasks helps to better understand the variations of how humans naturally structure and present information. In this respect, the results are an important basis for what a robot would have to deal with when it were in a learner's position. The tasks to be described are short and simple and are framed in such a way that a current robot – according to its vision and motor capabilities – would be able to perform the tasks. Moreover, the tasks were designed such that the teachers need to be fairly explicit in their descriptions and everyday knowledge is irrelevant for understanding the teacher's in-

structions. Both constraints are preconditions to make the information provided in the task scenario as self-contained as possible.

Participants: All in all, 22 people working or studying at universities in Munich participated in the data collection scenarios. Six out of these 22 participants explained and showed 2 of the four tasks to a robot. Although this sample of human-robot dyads is small, it already serves to gain first insights regarding differences in task descriptions when directed towards a human or a robot learner, see (Schreitter and Krenn, 2014).

Recorded were: the utterances, 3 videos – a frontal video of the teacher, a frontal video of the learner and a video of the setting, as well as motion and force data. In the current version of the corpus the audio and video data of the recordings are used for analysis and annotation, whereas the motion and force data have not been analysed and annotated yet. Overall, the data collection tasks resulted in 88 recordings comprising 12 human-robot (six in Task 3 and 4 respectively) and 76 human-human dyads. In 22 recordings the descriptions are directed towards the camera (Task 1), in 54 recordings the task descriptions are directed towards a human learner (22 in Task 2, 16 in Task 3, 16 in Task 4).
As not all teachers learned the tasks from participants, but from the experimenter, additional learners were required. Due to organisatory reasons five teachers explained the task to a 'knowing' learner, who was already acquainted with the task but instructed to act as if he/she did not know the task. In this study, the focus is on the information transmitted by the teacher. Although we are aware that 'knowing' learners react differently than naive learners, we argue that for our research questions it is sufficient that the teacher assumes that he/she is explaining the task to a naive learner.
In the following, the tasks, the reasons for construing the specific tasks and the setups for collecting the respective data are described. The focus in all tasks was on gathering multimodal information transmitted by the teacher, because this is information a robot should be able to process and analyse when confronted with task-oriented settings.

2.1. Task Scenarios

Task 1 (Figure 1)**:** Wooden fruits (a banana, a strawberry and a pear) are arranged and rearranged on a table. In this task, the teacher stands in front of a table and focuses on verbally explaining and manually conducting the task. There is no learner present. The items to be manipulated are a white sheet of paper on the left side of the teacher and a plate with three wooden fruits (a banana, a strawberry and a pear) on the right side, see Figure 1. Additionally, the teacher is equipped with a second sheet of paper depicting six steps of putting the fruits on certain locations at the paper and then reordering them. The teacher first describes the initial situation and then explains into the camera how to order the fruits from the plate on the white sheet of paper. One after the other, the three fruits are put on certain locations at the paper. Subsequently, two re-ordering movements of the fruits on the paper are conducted and the locations of two fruits changed.
This task was developed with a focus on auditory perception. All object names are voiced in order to produce audio recordings suitable for investigating auditory cues of information structure including prosody, givenness, and focus of attention.

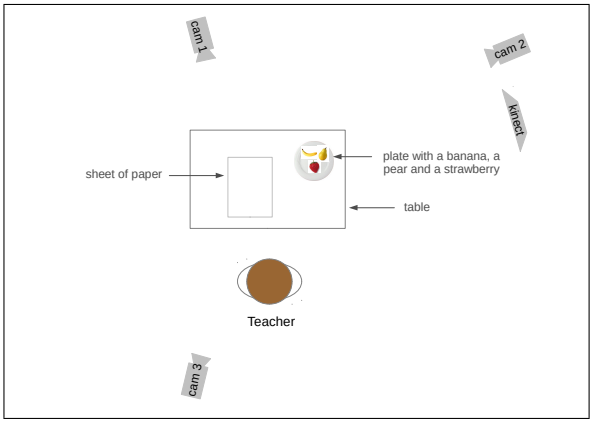

Figure 1: Task 1, arranging fruits (datasets from 22 humans)

Task 2 (Figure 2)**:** The goal is for the instructor and the learner to collaboratively move an object, standing at a table opposite of each other. On the table between the two participants, there is a board with two handles, see Figure 2. One handle is directed at the instructor and the other one at the learner. Both handles are marked with colours. When the task starts, the instructor asks the learner to grasp the handle at the learner's side with the left hand. The instructor grasps the handle at his/her side with the right hand. Then they lift the board and change position, i.e. they move around the table 180 degrees. Subsequently, they tilt the board 90 degrees, move along the table to the left side of the learner (i.e. the right side of the instructor), put the board down on the floor and lean it against the table.
For this task, the focus is on collaborative movement of a single object. In addition to explaining and conducting the task, the instructor has to observe whether the actions of the learner are correct.

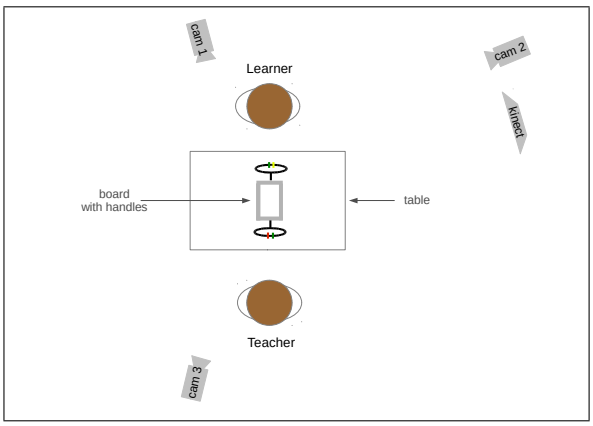

Figure 2: Task 2, collaboratively moving an object (datasets from 22 human-human pairs)

Task 3 (Figure 3)**:** A teacher explains and shows to a learner how to connect two separate parts of a tube and then to mount the tube in a box with holdings. The learner

stands in front of the table at the left side of the teacher (see Figure 3) and observes the task. Objects involved are a box with holdings placed on a table, a part of the tube already attached to the box and a loose part of the tube on an additional small table on the right side of the teacher. The loose part of the tube contains two coloured markers: a green and yellow one and a red and yellow one. 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.
The learner is only observing while the teacher is explaining and conducting the task. Therefore the learner has less influence on the task description than in Task 2.

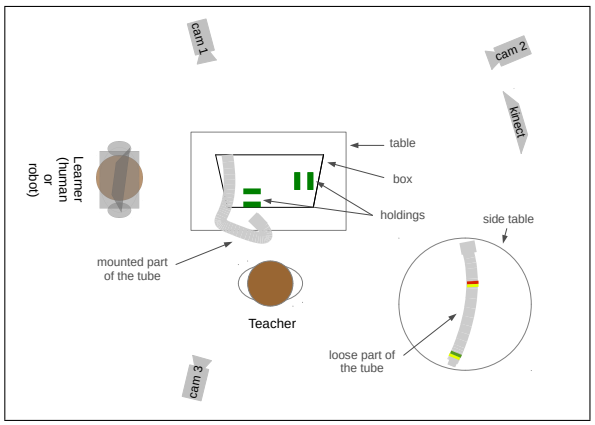

Figure 3: Task 3, mounting a tube (datasets from 16 human-human pairs and 6 human-robot pairs)

Task 4: (Figure 4)**:** The fourth task is a navigation task. The teacher instructs the learner which path to go to reach a chair. Inbuilt into the scenario is a path correction, where the instructor corrects and redirects the learner along a slightly different path.
In the room, there is a square table, a round table, a chair, and a small ball lying on the chair. Before the task starts, the learner is standing next to the square table, see Figure 4. The learner then has to pass the long side of the table, then the short side. Subsequently, the teacher asks the learner to walk around the round table towards the chair but does not say in which direction. The path on the left side and the path on the right side are equally long. When the learner initiates to move around the table in a certain direction, the teacher corrects him/her to walk around the table in the other direction. The learner then has to look at the chair and check if there is an object located on it. The teacher is explaining and the learner is conducting the task.

2.2. Human-Human and Human-Robot Dyads

Human-Human (HH) Dyads The first task presentation was directed towards a camera with the instruction that a person watching the video should be able to conduct the task. The second, third and fourth tasks were directed towards a human learner, who was told to carefully watch

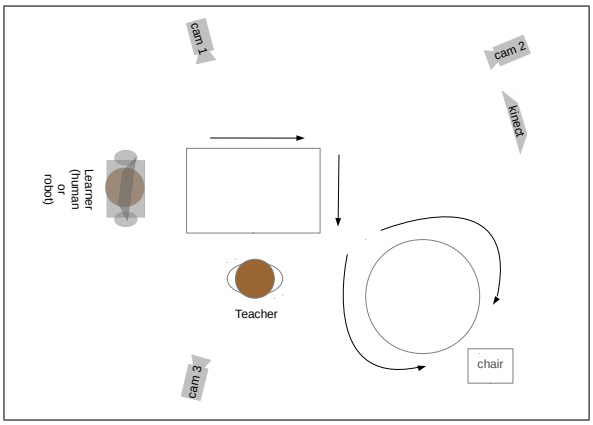

Figure 4: Task 4, navigation task (datasets from 16 human-human pairs and 6 human-robot pairs)

and listen to the explanations of the learner to be able to pass the information on to a new learner. In the subsequent trial, the learner became the new teacher. A calibration trial was introduced at least after every fifth trial where the experimenter functioned as teacher to counteract the Chinese whispers effect. The experimenter used the same wording each time. Additionally, before each task the teachers received a schematic 'cheat sheet' depicting the course of action during the task to reduce their cognitive load.

Human-Robot (HR) Dyads Teachers participating in the HR dyads explained the first task into the camera, the second task to a person and the third and fourth to a robot. They also received a 'cheat sheet' to reduce their cognitive load. The robot employed was a research prototype developed at the Institute of Automatic Control Engineering at the Technical University in Munich. It is of human-size height and is equipped with an omni-directional mobile platform, two anthropomorphic arms, and a pan-tilt unit on which Kinect sensors are mounted. Movement, head movements, and verbal feedback (e.g. *ja*, 'yes'; *ok*) were controlled by a human wizard. Empirical evidence has shown that non-verbal feedback from listeners such as eye gaze communicates understanding and is expected by human speakers (Eberhard et al., 1995). Additionally, speakers who do not get feedback from addressees take longer and make more elaborate references (Krauss and Weinheimer, 1966). Therefore we employed head-movements of the robot (so that the speaker was able to infer its eye gaze) and verbal backchannel feedback. The Kinect mounted on top of the robot (its 'head') was controlled by a Wizard-of-Oz during the task descriptions and directed either towards the setting or towards the face of the teacher. Additionally, the MARY Text-to-Speech Synthesis platform[1] was employed for giving verbal feedback during the task. For technical reasons, verbal feedback worked for five of the six participants.

Questionnaire In the HR dyads, the participants were additionally asked to fill in a questionnaire about their acquaintance with state-of-the-art in robotics and speech synthesis. This might influence their assessment of the robot

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

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

Only one out of the six participants had contact to robots before within a user study, and one with speech synthesis. Except for one, all participants had the impression, that the robot's head was controlled by an algorithm, whereas only two participants believed that the robot's navigation system was controlled by a computational algorithm, as opposed to four participants who believed that the robot was steered by a human. Overall, the naturalness of the interaction with the robot was 3.33 (SD: 1.21) on a five-point Likert scale (1 = very natural, 5 = not natural at all).

3. OFAI-MMTD Corpus – Data Analysis and Annotation

In the current version of the corpus (MMTD corpus V1 gold standard), annotated are the recordings of Task 2 and 3. This sub-corpus of 44 recordings is independently annotated by two annotators. The annotations have been merged and inconsistencies between the annotators were resolved.
Praat[2] was used for transcribing the utterances and annotating prosodic information. ELAN[3] was employed for the remaining manual annotations and for synchronising audio, video and representation tiers, thus, supporting analyses across modalities. In addition, Python programs were written to automatically extract temporal sequences of object references and respective cues on the different modalities.
The different layers of information annotated in MMTD corpus V1 gold standard are described in the following and sample annotations are presented.

3.1. Transcription and Transliteration of Teacher Utterances

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

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

The transcriptions were made in Praat for optimal temporal alignment of speech signal and transcription.

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

POS The transliterated utterances are then input to the TreeTagger (Schmid, 1995) and the thus resulting part-of-speech sequences are manually corrected. See line 3 of Table 1 for the annotations on the POS-tier. The labels stem from the Stuttgart-Tübingen Tagset[4].
The example in Figure 1 is taken from Task 3 where the teacher attaches the end of the tube with the red-yellow marker to the left green holding. Line 1 shows the trancribed utterance *und dann wos rot-gelb is*. (The full utterance is *und dann wos rot-gelb is in die Halterung* ('and then where it red-yellow is into the holding').) Line 2 shows the transliteration where *wos* is separated into *wo* and *es*. Line 3 shows the respective parts-of speech.

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

Table 1: Sample annotation: transcription-, transliteration- and POS-tier

3.2. Non-verbal Cues

Gesture of the teacher There exist a number of coding schemes for nonverbal behaviour, some of which are rather extensive such as the MUMIN ((Allwood et al., 2007)) and the BAP ((Dael et al., 2012)) coding schemes. The chosen coding scheme for gestures was adapted to the requirements of the corpus which comprises mainly object manipulation and deictic gestures. Thus, in the coding scheme deictic, iconic, beat, emblem and poising gestures produced by the teacher are manually annotated. In addition, for (i) deictic gestures, the object, location or person the gesture is directed at is annotated, for (ii) iconic gestures, the accordant action, for (iii) emblem gestures the kind of emblem that is used, (iv) for exhibiting gestures, the object emphasised by the gesture and for (v) poising gestures also the object emphasised by the gesture.
In the current version of the corpus, gestures are annotated along their category. If needed, further Elan tiers can be

[2] http://www.fon.hum.uva.nl/praat/
[3] https://tla.mpi.nl/tools/tla-tools/elan/
[4] http://www.ims.uni-stuttgart.de/forschung/ressourcen/-lexika/TagSets/stts-table.html

added with information on shape and movement dynamics stemming from the force and motion data which also were recorded as part of the data collection procedures.

Eye gaze of the teacher Where (to which object, location or person in the scenario) the teacher is looking to is manually annotated. Opposed to gestures, there is a continuous annotation of eye gaze over time.

3.3. Relevant Objects

On the "relevant objects"-tier the salient objects in the respective task description scene (excluding the learner/listener) are manually annotated.
For each task, a list of relevant objects is made. In Task 3 ("mounting a tube in a box with holdings"), for instance, the following objects are involved and, thus, need to be set into focus by the teacher for the learner to be able to follow the task. The respective objects are: a loose part of the tube, a mounted part of the tube, the two parts connected to one tube, a green and yellow marker, a red and yellow marker, green holdings at the right side of the teacher, and green holdings at the left side of the teacher.
On the "relevant objects"-tier the time span a specific object is salient is marked and the time span is labelled with the respective object label. In addition to the concrete objects involved in the task scenario, we have also foreseen a label for the task itself, as it is typical for the data in the MMTD corpus that the teachers refer to the task itself, typically at the beginning and the end of the task description.
The salience of an object is identified either by the occurrence of a linguistic reference in the teacher's speech, by the teacher's gaze behaviour, or specific communicative gestures such as deictic gestures, general communicative gestures (e.g., hands poising above objects in the field of attention), using fingers for counting, raising the index finger when talking about something important. Linguistic indicators are, for instance, full or elliptic noun phrases, e.g., *den Schlauch* ('the tube'), *Schlauch* ('tube'), pronouns or determiners, e.g., *er* ('it'), *der* ('the') for 'der Schlauch/the tube', determiners combined with deictic adverbs, e.g., *den hier* ('the one here'), space deictics, e.g., *hier, da* ('here', 'there'), adjectives, e.g., *rot-gelb* 'red-yellow' for the red-yellow marker attached to the tube. See the examples for salient objects in Tables 2 and 3, line 4. In the first example (Table 2), linguistic indicators for the salient object 'end of tube with red-yellow marker' are the deictic adverb *wo*, the personal pronoun *es* and the adjective *rot-gelb*. In Table 3, the salient object is 'the green holdings to the left of the teacher' co-occurring with the noun phrase *die Halterung*.

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

Table 2: Sample annotation: transcription-, transliteration-, POS- and 'salient object'-tier

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

Table 3: Sample annotation: transcription-, transliteration-, POS- and 'salient object'-tier

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

3.4. Prosodic Annotation

Prosodic information is annotated according to the DIMA annotation guidelines (Kügler et al., 2015). The DIMA approach has been chosen because it a) represents a consensus system for prosodic annotation of German; b) aims at compatibility of annotations and thus fosters the exchange of annotated data sets; c) allows for independent annotation of phrase boundaries, prominence levels and tones. As regards the MMTD corpus V1 gold standard, phrase boundaries and prominence levels are annotated:

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

Prominence level In a second annotation phase, prominent syllables are annotated with levels of perceived prominence. DIMA proposes three levels of prominence:

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

Note, in the MMTD data sets prominence levels 1 and 2 are predominant.
For an exhaustive presentation of the different tiers of prosodic DIMA annotation see (Kügler et al., 2015). Praat has been used for making the prosodic annotations. An annotation example from the MMTD data set is shown in Figure 5.

4. Early Use of the Corpus and Future Work

So far, the annotated data have been used to (i) suggest modifications which should be made to the Givenness Hierarchy (Gundel et al., 1993) in order to handle open world

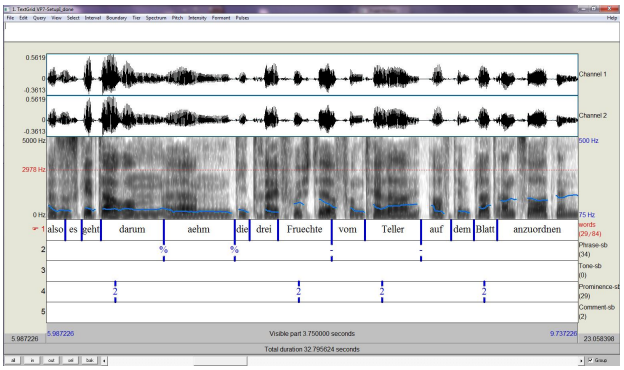

Figure 5: Sample annotation: phrase boundaries and prominence levels. % indicates strong phrase boundary, - indicates weak boundary, and 1 and 2 stand for prominence levels 1 and 2. The annotation example is taken from Task 1 data.

reference resolution (Williams et al., 2015), and (ii) to evaluate a computational model for situated open world reference resolution (Williams et al., 2016). A first detailed analysis of the multimodality of object references can be found in (Gross et al., accepted on 19 January 2016).
In the ongoing CHIST-ERA HLU research project ATLANTIS, the already collected data are further analysed and where necessary further annotated in order to model core competencies for multimodal communication in robots, enabling the robot

a) to draw attention to objects and their properties, and to spatial relations between objects: In a test environment a number of objects – where number, position, colour and type of objects vary – are scattered around the environment and robots have to use gesture, natural language (specific or underspecified) as well as other cues such as eye gaze and gesture, to draw attention to particular objects.

b) to talk about moving objects and guide robots and humans around an environment: In an environment of moving objects and robots, as well as (colourful) region on the ground and landmark objects robots shall be able to talk about paths of objects and guide other robots around the environment.

At the time of writing this article, the annotation of the tiers specified in Section 3. is ongoing for Task 1 and 4. Further tiers with information derived from recording force (Task 2) and motion data (Tasks 1-4) will be annotated. Moreover, the data from Task 1 will be annotated with a specific focus on information structure. In addition to the German dataset a comparable English dataset is available for Task 1, allowing us to compare the realisation of information structure in the German and English versions of the task descriptions.
The annotations will be made available to the research community, whereas the videos and sound files are subject to protection of data privacy.

5. Acknowledgements

The first author of the present paper is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Austrian Research Institute for Artificial Intelligence. The annotation work on the MMTD Corpus is in part funded by the CHIST-ERA HLU project "Artificial Language Understanding in Robots ATLANTIS". We gratefully thank Martine Grice, Stefan Baumann and Anna Bruggeman from the I*f*L Phonetik, University of Cologne for familiarising us with the DIMA guidelines for prosodic annotation, and our student co-worker Katharina Kranawetter for annotating parts of the corpus. 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.

6. Bibliographical References

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3-4):273–287.

Dael, N., Mortillaro, M., and Scherer, K. R. (2012). The body action and posture coding system (bap): Development and reliability. *Journal of Nonverbal Behavior*, 36(2):97–121.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., and Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of psycholinguistic research*, 24(6):409–436.

Gaspers, J., Panzner, M., Lemme, A., Cimiano, P., Rohlfing, K. J., and Wrede, S. (2014). A multimodal corpus for the evaluation of computational models for (grounded) language acquisition. In *Proc. of 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)@ EACL*, pages 30–37.

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307.

Krauss, R. M. and Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4(3):343.

Kügler, F., Smolibocki, B., Arnold, D., Baumann, S., Braun, B., Grice, M., Jannedy, S., Michalsky, J., Niebuhr, O., Peters, J., Ritter, S., Röhr, C. T., Schweitzer, A., Schweitzer, K., and Wagner, P. (2015). DIMA - Annotation Guidelines for German Intonation. In *Proceedings of the 18th International Congress of Phonetic Sciences*.

Nilsson Björkenstam, K. and Wirén, M. (2013). Multimodal annotation of parent-child interaction in a free-play setting. In *Thirteenth International Conference on Intelligent Virtual Agents (IVA 2013)*.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland.

Schreitter, S. and Krenn, B. (2014). Exploring inter- and intra-speaker variability in multi-modal task descriptions. In *Proceedings of the 17th IEEE International*

Symposium on the Robot and Human Interactive Communication (Ro-Man 2014), Edinburgh, Scotland.

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

Williams, T., Acharya, S., Schreitter, S., and Scheutz, M. (2016). Situated open world reference resolution for human-robot dialogue. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2016)*, Christchurch, New Zealand.

Yu, C., Smith, L. B., and Pereira, A. F. (2008). Grounding word learning in multimodal sensorimotor interaction. In *Proceedings of the 30th annual conference of the cognitive science society*, pages 1017–1022.