Phenomena in conveying information during oral task descriptions

Stephanie Schreitter and Brigitte Krenn

Austrian Research Institute for Artificial Intelligence, 1010 Vienna, Austria,
firstname.lastname@ofai.at

Abstract. A robot has to deal with a broad variety of information conveyed via verbal and non-verbal channels to be able to observe and listen to a task presented by a human teacher. We have collected a small corpus of human-human dyads to investigate how information is presented through verbal and/or visual channels. Apart from the characteristics of spoken language, the qualitative analysis of the data shows: (i) broad 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 attention of the learner, (iii) a temporal structuring of the task by verbal means for all teachers, and (iv) the use of generic "you" for most of the teachers.

Keywords: task descriptions, embodied language processing, oral communication

1 Background

In face-to-face communication, people do not only use speech but a multitude of non-verbal behaviours such as nods, communicative gestures, gazes, object manipulation gestures, etc. The vocal and the gestural acts together comprise the information necessary for the observer/learner to understand. Findings from embodied cognition have shown the importance of action and perception during language comprehension in humans [5, 6, 12, 13]. If robots are to interact with humans in natural ways in the future, a number of serious issues in multimodal communication must be tackled. With the present contribution, we aim at illustrating the problem and state (minimal) requirements for system functionality.

Imagine a robot that can analyse, interpret, and learn from task oriented presentations where a human teacher shows some activity to the robot learner and explains what she/he is doing by means of task accompanying speech. In the present paper, we investigate which kinds of communicative signals and their variations a robot should be able to deal with when it is presented with a task. We recorded human-human dyads to see which information is typically conveyed by which channels.

Clark and Krych [4], for instance, argue that human-human dialogue is a bilateral, opportunistic, and multimodal process where common ground is continuously updated. The authors emphasize 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. Gestures are an integral part of language, synchronous and co-expressive with speech, cf. [9, 1]. In a study by Lozano and Tversky [8], communicators explained how to assemble a simple object using either speech with gestures or only gestures. In the "gestures only"-condition, the assembly task was learned better and fewer assembly errors were made than in the "speech with gesture"-condition.

It is a challenge to equip robots with system components necessary to understand multimodal natural human communication. In a task description context, system components and the robot's architecture must (i) allow for robust incremental processing of natural speech and of multimodal communicative signals, (ii) include visual perception of the objects in the scene and the ongoing activity, and (iii) integrate all this in multimodal representations and the robot's episodic memory. Recent attempts have been made to address task-based natural language understanding on robots. Scheutz et al. [10] developed the robot architecture DIARC aiming at more natural human-robot interactions. The architecture includes mechanisms for natural language processing, intentional behaviours, and monitoring mechanisms to detect faults and recover from them. Kopp et al. [7] propose a model of how meaning can be organized and coordinated in speech and gesture. Their model is based on spreading activation within dynamically shaped multimodal memories.

The aim of the present work is to present and discuss an inventory of phenomena characteristic for showing and explaining to a learner what she/he should do. In the following sections, the corpus is presented and the communicative phenomena present in the corpus are discussed together with implications they have regarding specific components necessary for a robotic learner.

2 Corpus: human-human task descriptions

A corpus comprising 19 German recordings (video plus audio) was created where one person (the teacher or actor) showed another person (the observer or learner) how to mount a tube in a box with holdings, see Figure 1. [1] Two markers differing in colour had to be put in two different holdings. The teacher performed the task and verbally explained what had to be done. Thus, the task descriptions contain language mirroring the human perception and structuring of the task. The observer was told to carefully watch and listen to the explanations to be able to pass the information on to a new learner. The utterances were recorded as well as a frontal video of the setting including arms, hands, and torso of each teacher and learner. Although head and shoulders are not visible in the recordings, the transmission of non-verbal cues is already extensive. The manipulation task is borrowed from a robotic setting. Letting humans do the same task, and in addition let them explain it, helps to better understand what a robot would have to deal with when it were in the learner's position.

[1] The subjects were German students from the Technical University Munich (16 male, 3 female).

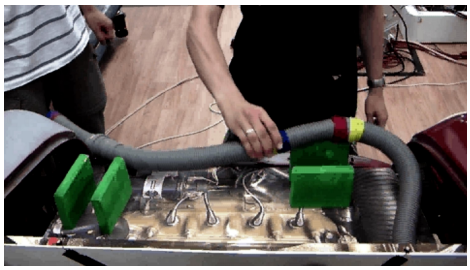

Fig. 1. A picture of the setting. A teacher is mounting a tube in a box with holdings.

3 Phenomena: how information is conveyed

The task to be described is quite simple: containing a grasp for the tube at a coloured marker, adjusting the tube between two green holdings, then grasping the tube at another coloured marker and putting it between another pair of green holdings. On average, the task duration was 21 seconds (12-34s). Although the task was quite simple and the learners had the assignment to listen carefully and forward the information to a new learner, there was quite some variation in how teachers presented the task. In the following, prevalent phenomena are presented and requirements for respective system components are briefly discussed.

Characteristics of spoken language Several properties typical for spoken language are present in the data: wrong word substitutions – 'holdings' (*Hindernis*)[2] instead of 'marker' (*Markierung*); repairs – 'red eh blue and yellow marker' (*rot äh blau-gelben Marker*); insertions – 'äh'; contractions – 'through the' (*durchs*, 'durch das'); errors – *habst* for 'have' ('hast').

These phenomena call for robust incremental language processing, e.g. [11], in addition to standard language technology tools such as automatic speech recognition, tokenization, part-of-speech, morphological analysis, phrase chunking, dependency parsing, and the such.

Variations in wording Objects in the task are the tube, two pairs of holdings, and three markers. For tube, all teachers used the same German word *Schlauch* (tube), except for three who did not verbally refer to the object at all. For 'marker', two teachers used the Anglicism *Marker*, and two used either 'point' (*Punkt*) and 'gripping point' (*Greifpunkt*) or 'endpoint' (*Punkt / Endpunkt*). The other 15 teachers used 'marker' (*Markierung*). For the holdings, there was a wide variation in naming: 'obstacle' (*Hindernis*), 'thing' (*Ding*), 'block' (*Block*), 'beam' (*Balken*), 'rail' (*Schiene*), 'marker' (*Markierung*), 'log' (*Klotz*), 'opening' (*Öffnung*). Again, there was one teacher who did not verbally refer to the holdings. The actions 'grasping the tube' and 'mounting the tube in the box

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

with the holdings' also showed some variance. For grasping, 'grasp' (*greifen*), 'have' (*haben*), 'take' (*nehmen*), 'span' (*umfassen*), 'change grip' (*umgreifen*) were used, and for 'putting the tube between the holdings': 'put' (*legen*), 'insert' (*führen / einführen / einspannen / einlegen / einsetzen / einfügen*), 'put inside' (*reinstellen / reinlegen*), 'clamp' (*klemmen*), and 'thread' (*einfädeln*).

Taking the above into account, the learner – may it be a human or a robot – has to infer objects and actions by listening and observing. The action is still the same, although 11 different verbs were used (up to two per teacher for the same action). Multimodal knowledge representations are a necessary prerequisite for dealing with lexical variation and omitted verbal references for objects and actions. The robot has to be able to resolve the connection of an abstract entity to an entity in the world, cf. [2], e.g. the words *Block*, *Klotz*, and *Hindernis* are all three referring to the green holdings. A comparison of what is visually perceived and what is uttered reveals how the same actions and objects are verbally expressed. In addition, the unspoken needs to be grounded in the scene. As Clark and Krych put it: "when the workspace is visible, the partners ground what they say not only with speech, but with gestures and other actions" [4], p.69. Thus, even though some teachers did not mention important elements of the task, the observers were able to understand.

Time markers Three teachers verbally signalled their respective learner that the task will now start, e.g. 'it is about' [...] (*es geht darum* [...]), 'the goal is' [...] (*Ziel ist* [...]), 10 told their learners when the task was done, e.g. [...] 'that was it' ([...] *das wars*), [...] 'that's all' ([...] *das ist alles*). All teachers used lexical time markers, such as 'first' (*zuerst*), 'then' (*dann*), 'subsequently' (*anschließend*) to signal the sequencing of the sub-tasks.

Therefore, as a technical basis, a (simple) model of before, after, and concurrency along a common timeline is required together with mechanisms to identify and interpret cues for temporal structuring. These may be lexical (as above), grammatical (tense) or determined by the course of multimodal action.

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

For a robot to be able to deal with these varieties, the following capabilities are required: (i) the ability to distinguish between the perception of self and

other, (ii) a robust interpretation of the perspective from which the action accompanying utterance is issued, and (iii) a model for taking initiative, i.e. for the observer to understand when to just go on observing and when to step in the actor's position.

Verbal and gestural references to visual perception 13 teachers verbally referred to objects, actions or locations, e.g. 'here' (*hier*), 'like this' (*so*), 'this obstacle' (*dieses Hindernis*). The most frequent kind of gestures during task explanations were deictic gestures and holds during object manipulation to refer to objects or actions in the visual scene. Both gestures serve as indicators for directing the attention of the listener to certain objects or actions. Three teachers used verbal references and communicative gestures simultaneously (e.g. 'here' (*hier*) + deictic gesture). One teacher neither used communicative gestures nor verbal references to the visual scene. He only mentioned the grasping of the marker and did not mention that the tube has to be mounted in the box with the holdings. This could only be inferred by the learner from the visual scene. In this respect, Herbert Clark argues that "placing things just in the right manner" ([3], p.243) is an indicative act in which an object is moved into the addressee's attention.

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

Verbal backchannels 10 learners signalled their understanding via verbal backchannels to their respective teacher, e.g. *ok*, *mhm*. Non-verbal backchannels such as head nods etc. were not visible in the present videos. The interplay of verbal and non-verbal backchannels in joint activity (speaking and listening together form a joint activity, cf. [4]) will be topic of further investigations.

4 Conclusion and future work

In this paper, we discussed phenomena occurring in a corpus of 19 simple task descriptions (action plus speech) of how to mount a tube in a box with holdings. They include characteristics of spoken language, variations in wording, verbal time marking, variation of teacher's perspective, and verbal and gestural references to the scene. These results highlight the importance of multimodal signal processing in human-robot interaction.

Depending on the phenomena and their functional challenge, there exist none up to a variety of proposed technical solutions for system functionalities. The interplay of the components and the requirement for real-time processing are still far from being reached. More research and integration work is needed on the way toward human-like task-based natural language processing for robots.

A second corpus has already been collected which is a follow-up and extension to the presented corpus. The new corpus includes 3 video streams – one of the teacher, the listener and the setup respectively – an audio stream, motion data, and force data when collaboratively manipulating an object. The interplay of head movement, eye gaze, gesture, facial expression, verbal and non-verbal backchannel feedback, body posture etc. will be further analysed based on the new data and the experiences from the initial corpus.

5 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 in recording the data.

References

1. Bergmann K.: The production of co-speech iconic gestures: empirical study and computational simulation with virtual agents. PhD Thesis. Bielefeld, University (2012)
2. Cantrell, R., Scheutz, M., Schermerhorn, P., Wu X.: Robust spoken instruction understanding for HRI. Proceedings of the 2010 HRI Conference. (2010)
3. Clark, H. H.: Pointing and placing. In S. Kita (ed.) Pointing. Where language, culture, and cognition meet. Hillsdale, NJ, Erlbaum (2003) 243–268
4. Clark, H. H., Krych, M. A.: Speaking while monitoring addressees for understanding. Journal of Memory and Language, **50(1)** (2004) 62–81
5. Horton, W. S., Rapp, N. D.: Occlusion and the accessibility of information in narrative comprehension. Psychonomic Bulletin & Review, **10** (2002) 104–109
6. Kiefer, M., Barsalou, L.: Grounding the human conceptual system in perception, action, and introspection. In Tutorials in action science. MIT Press (2011)
7. Kopp, S., Bergmann, K., Kahl, S.: 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)
8. Lozano, S., Tversky, B.: Communicative gestures facilitate problem solving for both communicators and recipients. Journal of Memory and Language. **55(1)** (2006) 47–63
9. McNeill, D.: Gesture and thought. University of Chicago Press, Chicago (2005)
10. Scheutz, M., Briggs, G., Cantrell, R., Krause, E., Williams, T., Veale, R.: Novel mechanisms for natural human-robot interactions in the DIARC architecture. Proceedings of AAAI workshop on intelligent robotic systems (2013)
11. Scheutz, M., Cantrell, R., Schermerhorn, P. Toward humanlike task-based dialogue processing for human robot interaction. AI magazine. **32(4)** (2011) 77–84
12. Stanfield, R. A., Zwaan, R. A.: The effect of orientation derived from verbal context on picture recognition. Psychological Science, **12** (2001) 153–156
13. Zwaan, R. A., Taylor, L.: Seeing, acting, understanding: motor resonance in language comprehension. Journal of Experimental Psychology, **135** (2006) 1–11