On Grounding Natural Kind Terms in Human-Robot Communication

Julia Peltason · Hannes Rieser · Sven Wachsmuth ·
Britta Wrede

Received: 15 November 2012 / Accepted: 8 March 2013 / Published online: 29 March 2013
© The Author(s) 2013. This article is published with open access at Springerlink.com

Abstract Our contribution situates Human-Robot Communication, especially the grounding of Natural Kind Terms, in the interface of Artificial Intelligence, Cognitive Psychology, Philosophy, Robotics and Semantics. We investigate whether a robot can be grounded in the sense favoured in Artificial Intelligence and Philosophy.

We thus extend the notion of grounding to social symbol grounding using an interactive perspective addressing the question how grounding can be achieved in detail in interaction. For the acquisition of Natural Kind Terms we establish the notions of foundational common ground and foundational grounding in contrast to the established common ground and grounding. We introduce the robot setting used and provide a deep evaluation of a tutorial dialogue between a user and the robot. We investigate these Human-Robot Communication data from an ethno-methodological and an “omniscient” perspective (the latter amounting to consideration of automatic speech recognition results) and test whether these perspectives matter for analysing grounding. We show that the robot has acquired a partial concept of a Natural Kind Term—represented by statistics over visual object features—and that this is shared knowledge, hence the first step of a grounding sequence. Finally, we argue that

grounding of robots can be achieved and extended to situated structures of considerable complexity.

Keywords Common ground · Grounding · Human-robot communication · Natural kind terms · Dialogue structure

1 Introduction

Grounding is a notion used in semantic and pragmatic theories dealing with the generation of (mostly) reliable collective or public information. The public involved may range from dyads to large communities. In the prototypical case of interest here, the information is the outcome of an information exchange by two or more conversational participants (CPs). The exchange may be varied as well; it may consist of speech acts or dialogue acts and sequences of these by speakers to addressees. In a simple case, the dialogue acts can be assertions of one CP to one or more addressees. If the addressees accept or acknowledge the information the speaker and the addressees share it, it is common to them. This common information can be presupposed in the following part of the dialogue or in new conversations which in a way rely on the current CPs. Common ground can also be produced by unknown sources and spread in the public by various media. The dyad we will deal with consists of a robot and a human instructing the robot. This is an entirely new variant in the grounding domain. In human-human communication (HHC), the obstacles to grounding and the production of common ground are numerous. They range from problems of hearing to disbelieved information and arguments regarded with suspicion. Every type of obstacle comes with its own means for remedy. Hearing problems, for example, can be treated with clarification

J. Peltason · B. Wrede (✉)
Applied Informatics, Faculty of Technology, Bielefeld University,
Bielefeld, Germany
e-mail: bwrede@techfak.uni-bielefeld.de

H. Rieser
Faculty of Linguistics and Literary Studies, Bielefeld University,
Bielefeld, Germany

S. Wachsmuth
Center of Excellence, Cognitive Interaction Technology (CITEC),
Bielefeld University, Bielefeld, Germany

requests, disbelieved information by denial and suspect arguments by counterarguments. In our paper we start from this well-established notion of common ground and investigate whether it can be a guiding concept in robotics, more precisely, in human-robot communication (HRC).

In contrast to HHC, HRC is highly asymmetric in terms of capabilities, background knowledge, and perception. The robot senses the world through image pixels that are grouped to coarsely characterised visual blobs. It understands speech based on a limited vocabulary that is matched to any sound input—possibly causing speech recognition errors. Nevertheless, both interlocutors—the human and the robot—are situated in the same physical world and are able to refer to the same real world objects. Thus, a partial overlap of their mental models is practically possible raising the question whether the robot’s capabilities are sufficient to achieve a common ground in communication.

Word meaning must first and foremost be grounded, i.e. made socially available. We take as our paradigm case the grounding of natural kind terms (NKTs) for the robot through a user. Only if the words are grounded, can the proposition made up by them be grounded as well. So, intuitively, the grounding business for robots will start at the level of words. We defend this idea in the next section. In our modelling of robotic interaction we favour an approach which is not explicitly mentalistic but descriptive in an externalistic sense, focusing on agents’ contributions. As a consequence, grounding is not explicitly modelled by an algorithmic procedure. Nevertheless we are interested in grounding procedures. The reason is that the robot’s command of an NKT should be backed by common knowledge, i.e. socially anchored. If this is the case, the robot uses the NKT according to accepted lexical conventions. Therefore, we must investigate, how much of a classical grounding concept can be satisfied by the robot-user interaction implemented. If a tutorial dialogue introducing a NKT is successful, i.e. if the robot agreed to memorise the NKT in question and this in turn is acknowledged by the user, we have evidence that our system reconstructs acquisition of NKTs up to revision as in human-human dialogue.

The paper is structured as follows. In Sect. 2 we briefly introduce the formal concept of common ground and discuss how it relates to our HRC-setting. Section 3 describes the standard grounding process applying Clark’s Action Ladder and points out the need of a foundational grounding process in the usage of words. Section 4 presents the tutorial setting for the acquisition of Natural Kind Terms by the robot, the system capabilities needed, and the experimental setup. An example dialogue of this dataset is discussed in Sect. 5 which is analysed from an ethno-methodological perspective and an omniscient one. We draw a conclusion of what has been acquired by the robot in Sect. 6 and summarise our findings in Sect. 7.

2 Collective Mental States and Dialogue

In this paper we go beyond the concept of a grounding process for an individual system. Instead, we consider how symbol grounding can be achieved in a step-wise process between two interaction participants. We also have grounding of word forms which, however, we do not treat in this paper. Collective or common information is a necessary ingredient of our everyday doings exploiting coordination and cooperation of agents [21, 22]. It makes up the essence of some fact being grounded, i.e. of some fact being collective or common information. Grounding is achieved using a set of procedures aiming at establishing common ground. Frequently, grounding is implemented using types of dialogue acts [19] which themselves also had to be grounded at some time, a fact crucial for HRC as we will see. The role of dialogue in grounding has been a prominent research topic of Clark and colleagues, see [5, 7]. How can the concept of common information be explained? To sum up a discussion which began with [13] and has been continued since then, common information is circular or infinitely recursive [1, 2]; it does therefore not admit of a formulation as a finite iteration of mental states attributed by one agent to another and *vice versa* in the manner of *I know that you know that I know* and so on for n times and *You know that I know that you know* and so on for n times. The non-finiteness argument is due to [6] and has been accepted since then. Perhaps the most elegant formulation of the circularity property, referred to as (CIRC-Def) below, was given by the philosopher Harman [11] using mutual belief as a proxy for common information:

(CIRC-Def) a and b mutually believe that $p := (q)$ a and b believe that p and that q .

Note that in (CIRC-Def) “ q ” points to “ (q) ” establishing the circle. (CIRC-Def) does not contain features of some anchoring situation. As a contrasting example of a so-called shared situation definition (SH-SIT-Def) we present here Clark’s neo-Lewisian version [4] (p. 66), where the situation is taken explicitly into account. (We consider the relation between the (SH-SIT-Def) and our human-robot-setting below):

(SH-SIT-Def) p is common ground for members of community C iff $:=$

1. Every member of C has information that basis b holds;
2. b indicates to every member of C that every member of C has information that b holds;
3. b indicates to members of C that p .

In the (SH-SIT-Def) circularity rests on the use of “ b ”. Clark generalizes the notion of Common Belief/Knowledge by taking “having information” as a cover term for believing, knowing, being aware that, supposing that and factual

versions of seeing.¹ Anything able to carry information can serve as the basis b for C , especially things seen and utterances. We will use these properties in Sects. 5 and 6. (SH-SIT-Def) captures a notion which is useful in most empirical contexts, since it allows to anchor having information that p to situations and groups C ranging from dyads onwards.²

We emphasize that all interpersonal routines that speakers of a language *qua* members of a culture are adjusted to depend on common ground in this sense. Hence information (that) p covers word meaning as well as rules of turn taking in dialogue and much else. “Much else” comprises e.g. the opening, the task-related part and the closing of the HRC tutorial dialogue used in Sects. 3 and 4 as well as the tutorial dialogue itself taken as a structured entity. Since grounding has to be ultimately anchored in individual mental states, it is subjective and public or social at the same time.

By way of an interim sum up, (SH-SIT-Def) is mapped unto our HRC-setting in the following way: The situation b we are interested in is the tutorial situation between the robot and a user demonstrating a pineapple to it and introducing its NKT “*pineapple*”. C is the dyad of user and robot and p is the information *pineapple(demonstrated object)* or *pineapple(this)*.

3 A Working Concept: Clark’s Action Ladder and Types of Common Ground

If grounding is ubiquitous as stated above, how are communication and grounding related? This is one of the central issues of our paper. There are various answers to that and the one we will subscribe to depends on Clark’s Action Ladder for grounding. Clark’s Action Ladder describes a hierarchy of four levels which must be traversed in order to anchor the speaker’s meaning of a proposition p for an addressee. For reasons which will become clear soon, we refer to Clark’s notion of common ground as standard common ground (STCG) and the procedures used for grounding as standard grounding (STG). Summarising, the Clarkian pathway to STCG is as follows: Dialogue is a form of coordinated action. Coordinated actions are joint projects. These are, respectively, proposed and considered by the conversational participants (CPs). To guarantee the success of joint projects one needs STG and STCG. Here is an example of STG-ing a word [4] (p. 221):

Roger: now, - um do you and your husband have a j- car
 Nina: - have a car?
 Roger: yeah
 Nina: no-

¹Not all of these are veridical, so the generalization has non-trivial consequences.

²AI research in modal logics uses concepts like (CIRC-Def) and (SH-SIT-Def) as well, see the axiomatic accounts in [8] and [15] for comparison.

Note the structure of the STG sequence: A problem is identified by Nina and treated using a clarification request, attended to in turn by Roger. We get a repair from “*j-car*” to “*car*” which is accepted; then the base-line conversation continues with Nina’s answer. What we see in the example is that a problem arises and is solved by a side sequence (“- *have a car?*” “*yeah*”). The solution yields an update: Its result can be added to the STCG accessible to both CPs. What goes into the STCG can be seen from Clark’s Action Ladder³ (which corresponds to a systematics from bottom to top)⁴ as shown in Table 1.

The level interaction represented in Table 1 is: A presents X and B accepts X and provides [positive] evidence in the form of feedback. STCG must exist for all levels. STCG on higher levels presupposes STCG on lower levels. The “Principle of joint closure” regulates presentation and acceptance: “The participants in a joint action try to establish the mutual belief that they have succeeded well enough for current purposes”⁵ (p. 226). The STCG is updated on all of these presentation + acceptance levels. Given [belief in] successful updates, that is if there is no need for clarification, the dialogue can move on. All layers, especially three and four, can lead to production of adjacency pairs. More intricate dialogue structures can be modelled based on these and other projective patterns, for example genre regularities as in tutorial dialogue. The link between signalling meanings p felicitously and grounding aiming at STCG is that every level of the hierarchy must be standardly grounded, for example identification of words and understanding of word meanings finally making up p . In short, every level is “mutually approved” as satisfactory by the CPs. Now an important difference between the example above and the HRC becomes visible. STCG does not deal with the acquisition of terms. In the example above the meaning of the concept car is not at issue, it is presupposed. In contrast, in our HRC setting, the meaning of natural kind terms (NKTs) has to be grounded (as have all other meanings), if communication is to be natural and successful. In a way, the grounding problems we face in HRC are nearer to the acquisition of meanings. We call these types of common ground and grounding foundational common ground (FCG) and foundational grounding (FG). Our claim is: If a robot should be enabled to communicate, its usage of words has to be foundationally grounded. Grounding words is a precondition for the robot’s behaving socially, and *vice versa*. We look at a tutorial dialogue between a human and a robot to investigate whether this is a

³Meanwhile there are more elaborate versions of the Action Ladder. However, for our purposes the grain of the original one suffices.

⁴Clark and most other theorists of STCG share a preference for relying on “that p ”, i.e. on propositions but other options are necessary in the end.

⁵Note that STCG is defined via mutual belief here.

Table 1 Clark's action ladder [4, p. 221]

Speaker A's actions/presentations	Addressee B's actions/acceptance
A is proposing joint project w to B	B is considering A's proposal of w
A is signalling that p for B	B is recognizing that p from A
A is presenting signal s to B	B is identifying signal s from A
A is executing behaviour t for B	B is attending to behaviour t from A

feasible claim, empirically and methodologically. In passing, we already pointed out how the grounding concept applies to the dialogue genre of “tutorial dialogue”, conceived of as a series of joint projects, which is central for our instance of HRC. This will be initialized with a description of the robot's dialogue interaction patterns in Sects. 4 and 5. Concerning NKTs a few remarks must suffice (see [3] for more information). They make up our vocabulary for classifying naturally given objects, examples are names for fruits (our type of example), liquids, trees, flowers, animals, body parts and so on. They are tied to perceptual experience and admit of simple hierarchies. They are basic from the developmental and the evolutionary point of view and universal, i.e. they can be found in all languages.

4 The Robot Setting Used: “Curious Flobi”

In this section we describe the facilities for acquisition of NKTs by the robot used, especially the interaction of the system with a Wizard-of-Oz (WOz) component. Observe that the semantics of NKTs is based on visual perception. This implies that the robot has to integrate his percept of the fruit indicated into the meaning of the NKT at stake. This fact motivated the robotic setting used (deixis, WOz) in the first place.

4.1 The Underlying Dialogue Model

In linguistic dialogue modelling, roughly two main traditions can be identified: approaches that model the internal attitudes of the interaction partners and their underlying cognitive state and, in contrast, approaches that describe the public and conventional aspects of how interaction typically proceeds [9]. Accordingly, also approaches to computational dialogue modelling can be categorized into *mental-state approaches* and *descriptive approaches*. Descriptive approaches include all types of models that specify the dialogue flow explicitly, while in mental-state approaches the dialogue flow is created dynamically, emerging from a model of the interaction goals or of the interaction partner's mental state.

Despite much of research carried out in both directions, interactive robotic systems bring new challenges to dialogue

modelling. It has now become necessary to take physical situatedness into account. This means that questions of reactivity to dynamic environments, possibly involving multiple modalities, and of potentially open-ended, unstructured interactions, involving multiple tasks at a time play an important role and have to be considered in the dialogue model. However, existing approaches to dialogue modelling—both mental-state and descriptive—have been focusing on designing interactions for information-oriented query systems and are thus not directly transferable to robotics, as argued in [17]. We have therefore suggested a new approach to dialogue modelling that considers these special demands of robotic applications [16], called PaMini (Pattern-Based Mixed-Initiative Interaction), which we used to implement the tutoring dialogue with the robot Flobi.

The PaMini approach relies on a set of generic *Interaction Patterns* that capture recurring dialogue structures, such as asking for information or requesting an action. Figure 1 shows the interaction pattern used for information requests (initiated by the robot) and object demonstrations (initiated by the human) as an example. At run-time, several such patterns can be active at the same time, and can be interleaved to achieve a more flexible interaction style. As an interface to the domain processing of the robot system, PaMini makes use of a fine-grained *Task State Protocol*: a task gets initiated, accepted, cancelled or updated, may deliver intermediate results and is completed finally; alternatively, it may be rejected by the handling component or its execution may fail. The dialogue manager is notified about task state changes and reacts by generating appropriate utterances. By combining these task states with robot dialogue acts, the above Interaction Patterns relate the conversation level with the domain level and integrate the robot's action and perception into the dialogue. A major difference to non-situated approaches to dialogue modelling is that the dialogue flow (i.e. the sequence of activated Interaction Patterns) is not decided by the dialogue manager alone, but decided externally in reaction to real-world events.

PaMini can be referred to as a descriptive approach, but it operates at a more abstract level of dialogue than traditional descriptive approaches, thus enabling reusability of dialogue strategies, and with its interleaving patterns it allows for a more flexible flow of dialogue. In contrast to many mental-state approaches, such as Traum's computational theory of

Fig. 1 The Interaction Patterns used for information requests and for object demonstrations, respectively

grounding [20], the approach of Interaction Patterns does not model the grounding process explicitly. Rather, grounding is incorporated implicitly within the Interaction Patterns. On successful completion of e.g. an object demonstration or information request pattern, the negotiated information can be viewed as shared knowledge, but not as grounded in the standard sense, as we will see in Sect. 6. Using the Task State Protocol, it is passed to the robotic subsystem, where it will be stored or processed further.

4.2 System Capabilities

In the Curious Flobi tutoring scenario, shown in Fig. 2, a number of previously unknown objects are present on a table. The robot acquires their labels, i.e. the NKTs, through spoken natural language dialogue with a human tutor. The human can show objects to the robot, label them, and verify what has been learned. Following a mixed-initiative interaction style, the robot can also ask for object labels on its own initiative.

The dialogue strategies of the Flobi system were designed based on a Wizard-of-Oz study on object teaching, where uninstructed users demonstrate everyday objects to a teleoperated robot system [12]. All of the analysed interactions have a very similar structure, consisting of an opening part, a task-related part and a closing part. 36 % of the interactions additionally feature transitional phrases that introduce the task-related part. In the opening phase, introducing each other (82 %) and exchanging pleasantries (18 %) are frequent. Aside from object demonstrations, the task-related phase consists of checking learned objects (45 %)

Fig. 2 Scenario overview

and transitional phrases between the objects (36 %). Praising the robot for correctly learned objects turned out to be universal (100 %). The task-related part may include closing remarks (36 %). As far as possible, the observed strategies were transferred to the Flobi scenario. Table 2 illustrates the resulting interaction capabilities of the system.

4.3 System Overview

Our system uses a component-based architecture with event-based, asynchronous communication. Figure 3 shows the components of the system, consisting of three subparts for vision, speech and motion. The major components involved in object demonstration episodes (on which we will focus for the purpose of this paper) are marked in orange colour. As shown in Fig. 1, processing an object demonstration requires two sequential steps: First, the object referred to needs to be identified (i.e. a *reference resolution* task needs to be executed). Second, its appearance must be memorized

Table 2 Example dialogue demonstrating the interaction capabilities of the system

Purpose	Example interaction
Greeting	User: Hello, Flobi Flobi: Hi, hello
Introducing	User: My name is Anneliese Flobi: Hello, Anneliese. Nice to meet you
Exchanging pleasantries	User: How are you? Flobi: Fine, thank you!
Task transition	User: I am here to show you some objects Flobi: Alright, let's start!
Attracting attention	User: Flobi, look here! Flobi: Yes, I'm looking
Object demonstration	User: This is a yellow lemon Flobi: Alright, a yellow melon. Is that correct? User: No, it's a lemon Flobi: Alright, a lemon. Is that correct? User: Yes, that's correct Flobi: Okay, I am going to learn the lemon Flobi: I have learned the lemon
Object query	Flobi: What is that? User: This is an apple (remainder similar to object demonstration)
Listing learned objects	User: What have you learned so far? Flobi: I have learned a lemon and an apple
Checking	User: Flobi, do you know this one? Flobi: I will have a look at it Flobi: This is an apple
Praising	User: Well done, Flobi Flobi: Thank you
Restart	User: Restart Flobi: Okay, let's start over again
Transitional phrases	User: Would you like to learn another object? Flobi: Yes, I want to learn them all
Closing task	User: Okay, that's it Flobi: Are you leaving already?
Parting	User: Good bye, Flobi. Take care! Flobi: Bye, see you later

(i.e. an *object learning* task needs to be executed). Both tasks are initiated by the dialogue system, and handled by a reference resolution and an object learning component, respectively. The components involved coordinate by means of the above described Task State Protocol.

The system is not (yet) fully autonomous. In particular, the reference resolution is operator-assisted using the Wizard-of-Oz method. This is because, despite our general goal of using only autonomous behaviour, we found through the above mentioned study on object teaching that users' referencing behaviour varies considerably, and this observation was confirmed by the user study with the Flobi sys-

tem, where 8 different referencing strategies could be identified. Not all of these could be acceptably automated by now. Hence, the system detects references on its own, but resolving is done by an operator selecting the appropriate object region. Thus, the normal flow of events when processing an object demonstration is as follows (cf. also Fig. 1):

- The user produces an utterance. The dialogue system interprets the user's utterance as an object demonstration (H.demonstrate, e.g. "U: This is a pineapple.").
- The dialogue system initiates a *reference resolution* task.

Fig. 3 Components of the Flobi system. The major components required for object demonstrations are: Speech recognition, dialogue manager, reference resolution and object recognition. They are marked in orange colour (Colour figure online)

- The WOz reference resolution component captures the current image of the scene so that the operator can select the object referred to.
- The WOz reference resolution component updates the *reference resolution* task with the coordinates of the object and completes it.
- The dialogue system generates a request for confirmation of the given NKT (R.askForConfirmation, e.g. “This is a pineapple, is that correct?”)
- If the user confirms (H.confirm, e.g. “Yes, that is correct.”), the dialogue system generates an acknowledgement (R.acknowledge, e.g. “Good.”) and initiates an *object learning* task, using the coordinates obtained from the *reference resolution* task.
- If the object recogniser accepts the learning task, the dialogue manager announces the start of the learning process (R.assert, e.g. “R: Then I will learn the pineapple now.”)
- The object recogniser updates its representations and completes the learning task which the dialogue system acknowledges (R.acknowledge, “R: I have memorized the pineapple.”).

Note that the operator who resolves the reference does not have complete information, but observes only the (possibly incorrect) speech recognition result as well as the user’s deictic gesture through the robot’s eye cameras. Thus, it relies on the same information an autonomous component would have to. This means that reference resolution may fail, which occurs for instance if the deictic gesture is performed out of the robot’s field of view or if the user does not perform a deictic gesture at all but refers to the object only verbally (“There is an apple next to the banana”).

4.4 The Dataset

The dialogue excerpt we will analyse in this paper is taken from a study with unexperienced users [18]. For the study 32 participants aged between 21 and 79 interacted with the Flobi system, with little prior instruction. Participants were told that the study was about object learning and asked to fill in a pre-questionnaire that captured their expectations towards the (still inactive) robot. Subsequently, they were asked to interact with the robot for at least 10 minutes and informed that they could begin interaction by greeting the robot and end interaction by saying goodbye.

In addition, they were advised not to be discouraged by speech recognition problems, and an emergency phrase (“Restart”) was provided. In order to obtain natural demonstration behaviour, however, it was not specified how exactly they should present the objects to the robot. Following the interaction, participants were asked to fill in a closing questionnaire which contained both statements about the interaction and follow-up judgement questions, matched to the expectation questionnaire.

A wide range of *objective measures* from the categories dialogue efficiency, dialogue quality and task success was collected, most of them obtained automatically from system log files. The questionnaires captured *subjective measurements* by the participants, including items that refer both to the interaction (dialogue efficiency, task success, cooperativeness and usability) and to their impression of the robot (likability, perceived intelligence, animacy, task abilities, personality and predictability). In order to determine the relevant factors that contribute to the various aspects of user satisfaction, the objective and subjective measures were related with each other as proposed by the PAR-

ADISE method [23]. In addition, we varied the degree of the robot's task initiative as a three-level between-subjects factor. These various quantitative evaluations were complemented by qualitative analyses of selected interaction phenomena, from which we will present a sample in this paper.

5 The Tutorial Dialogue Introducing the NKT “Pineapple” in a HRC Session

In this section we present a tutorial dialogue between a human and the robot which finally leads to the acquisition of the NKT “pineapple”. We scrutinize the public contributions of the agents as well as the automatic speech recognition (ASR) decodings of the robotic system. The dialogue is evaluated relying on principles of CA, modern dialogue theory, theories of intention and cooperation and much else of current research in communication (see bibliography). A special focus is placed on the nature of the NKT pineapple acquired. How much of it is acquired by the robot? We walk through the Action Ladder hierarchy and specify what we have got in the HRC at stake. At first sight, this seems remarkably little but the first sight deludes. We treat this more closely in Sect. 6. First we start in Sect. 5.1 with the dialogue (Fig. 4) and a technical description of what happens from the point of view of the “omniscient engineer”.

5.1 The Technical Perspective

For the understanding of the dialogue excerpt shown in Fig. 4, it is helpful to know that utterances are interpreted based on a speech recognition grammar. In more detail, the automatic speech recogniser (ASR) generates for each utterance one or several grammar trees representing (parts of) it. The dialogue system (DLG) then uses the nonterminals of the grammar (e.g. Greeting, ObjectDescription) as semantic tags. Based on conditions on them, the appropriate interaction pattern is triggered. For instance, the *object demonstration* pattern is triggered if the parse tree contains an *ObjectDescription* nonterminal, and a *greeting* is triggered if the parse tree contains a *Greeting* nonterminal but no task-related nonterminal. Furthermore, pattern selection is context-sensitive: the dialogue system's expectations influence the order in which the conditions are tested.

The handling of the dialogue shown in Fig. 4 is as follows:

- The user's utterance (81.) is split into two parts, due to the user making a short pause. As we will see in the following, the surplus utterance produced by the split-up disturbs the correct association of robot and user utterances. For the user, however, this is not evident.

- First, the DLG processes the first part of the parse tree (“until soon a pineapple”). Despite ASR misclassifications, the DLG interprets it correctly as an object demonstration, triggering an *object demonstration* pattern and leading to the robot's asking for confirmation of the label.
- The user in fact confirms the robot's request (83.), but the DLG needs to process the second part of the split-up utterance first (“upon the table”). As this does not constitute a valid reply to the confirmation request, the robot repeats its request (84.).
- The user confirms a second time (85.). However, the user's first confirmation (83.) is associated with the second request (84.). The second confirmation remains unprocessed.
- Having received the confirmation of the label, the DLG starts the learning of the label (86., 87.) and acknowledges if completed (88.).
- The user praises the robot (89.) to which the robot replies by thanking the user (90.). Even though the robot's reaction seems coherent, it is in fact the reply to the user's still unprocessed second label confirmation (85.) which the DLG interprets as praising because no confirmation is expected in the current dialogue context.
- Finally, the user's actual praising (89.) still needs to be processed. Misclassified as a greeting, it causes the robot to greet back (90.), which explains the robot's unmotivated greeting at the end of the tutorial dialogue.

5.2 Two Perspectives, Ethnomethodological Versus Omniscient

We first turn to the “ethno-methodology of HRC” and take the CPs' contributions as turns in a normal dialogue. As a consequence, we do not integrate ASR data, since in normal communication we also do not have access to internal representations. From this perspective, we have the user's demonstration of the pineapple in (81.); the robot acknowledges and asks for acknowledgement in (82.), given in (83.). From a CA or a dialogue perspective we are then done with the “question under discussion” [10] acquiring the NKT pineapple and the CPs could either move on to a new discourse topic or terminate the encounter by a closing sequence (see Fig. 1). However, we get again an acknowledgement and a request for acknowledgement without any indication of a clarification request in (84.). The user politely acknowledges (85.), the robot accepts (86.). He indicates memorization of the concept pineapple (87.) and successful completion (88.), again accepted by the user (89.). We get the robot's thanks (90.) and his unmotivated greeting (91.) which is a flaw, interactionally speaking. However, the external datum looks quite acceptable from the point of view of ethno-methodology and dialogue theory, intentions' analysis would perhaps falter over (84.) and (85.).

the latter iteration being justified by the user's acknowledgement in (89.). What we do not get however, is $Know_{robot} Know_{user}(pineapple(demonstrated\ object))$, i.e. the robot knowing that the user knows that the demonstrated object is a pineapple, due to a lack of inferential capability on the robot's side. As a consequence, neither (CIRC-Def) nor (SH-SIT-Def) can be satisfied. Turning to the omniscient point of view, ASR information seems to indicate that we do not get at the propositional level in the Action Ladder, and that hence communication is not achieved, the "C" in HRC threatening to be a misnomer for our datum. A closer look reveals, however, that the multi-modal setting yields propositional content. Propositional information, being more abstract than merely verbal information, is not exclusively tied to verbal expressions. And we have the reference resolution induced by WOz. Therefore we arrive at $pineapple(demonstrated\ object)$ as well. Granted that we get this proposition, what is its status wrt grounding? Well, the reasons for (Shared-KnowI) are the same as in the ethno-methodological case. Using an iterative account of grounding just for the moment, we can argue that we have reached the first stage in the "grounding hierarchy", namely, the agents' shared knowledge that the demonstrated object is a pineapple, trivial for the human but not so trivial for the robot. Anyway, we have already arrived at a socially active concept. The user might reason (erroneously, as we know, overestimating the robot's capabilities):⁶ Now the robot knows what a pineapple is, as good as I know, and after my acceptance he knows that I know. We have seen on which properties of the tutorial dialogue the acquisition of the NKT structurally depends. On the surface, it is the joint project NKT-introduction between the human and the robot. Looking deeper, however, we realize that the schema for the joint project is itself not well grounded: The robot does not KNOW turn taking rules, so it cannot project (anticipate) sequences in the CA sense. Certainly, grounding turn taking rules would be a feasible but demanding HRC project.⁷

⁶It is certainly true, as one reviewer indicated, that the user's terms and the robot's terms are of different semantic depth. So the question arises, whether the robot's answers should have clearly indicated this difference, so instead of "learn the pineapple" it should perhaps have used something like the more appropriate "I can visually recognise the image of a pineapple". In our study, we tried to reach an overlap between the user's semantics of NKTs and Flobi's newly acquired NKT. We were eager to keep the rest of the terms used by the robot as simple as possible, thus avoiding additional grounding problems. Our claim is, then, that there is a substantial intersection between the semantics of the NKT of the user and the robot after the positive conclusion of the tutorial dialogue. "Learning" in the robot's dialect has the purely extensional sense of establishing a stable association between the word form and the visual shape indicated by the user's deixis. To do the same for "visually recognise" and "image" would have been much harder. Granted, that this falls short of the user's meaning for "learning" which is determined by the user's biography.

⁷In addition, the robot does not develop Searlean intentions which are also a precondition for public communication as well as Gricean inten-

7 Conclusion: Robot Mentalism

We started from the now classical concept of grounding based on definitions given by D. Lewis, G. Harman, and H.H. Clark. Trying to apply it to HRC, we found that the classical concept, oriented towards HHC, presupposes a lot, above all that the CPs under investigation share the same language, especially its semantics. This is not the case for robots' instruction. There, grounding has to be built up from the very bottom. This is exactly what happens in the tutorial dialogue between the user and the robot acquiring an NKT, where the NKT consists of a word form like "pineapple" and the perceptual image of the fruit as introduced by the user's deixis. This insight led us to distinguish between standard grounding (STG), standard common ground (STCG)—the received point of view—on the one hand and foundational common ground (FCG) and foundational grounding (FG) on the other hand. FG and FCG anchor concepts in the world. Clark's hypothesis is that in understanding propositions every level of the linguistic hierarchy involved has to be fully grounded (has to be in STCG). At least for our robotic setting this claim is too strong. We demonstrated that there can be short-cuts: Through unification of the user's deixis and the reference resolution of WOz a proposition like $pineapple(demonstrated\ object)$ is generated which is also KNOWN by the user and the robot, due to world knowledge on the user's side and the mechanics of the tutorial dialogue on the robot's side, cf. the robot's "I have myself the pineapple memorized". Since the user and the robot KNOW the proposition, they have SOME shared knowledge and partially satisfy the shared situation definition (cf. SH-SIT-Def), the anchor of the ascending hierarchy of alternating knowledge operators. How do we finally assess the status of the robot's mental state?

On the whole we support the case of robot mentalism, but we stress that common ground for selected issues—here the foundational grounding of NKTs in a tutorial dialogue—can be achieved satisfying the (SH-SIT-Def) without establishing STCG on all levels of the action ladder. Our study shows that we can safely argue against sceptics which *in toto* want to deny the status of being grounded to a robot. The tutorial dialogue arrives at a socially active concept for pineapple

tion recognition. This means that the robot does not attribute intentions to the user, for example "he wants me to identify the pineapple now and to memorise it". Recognising the intention of the other is important for understanding speech acts and reacting accordingly, e.g. provide partial acknowledgments. Nevertheless, we have achieved knowledge of a NKT for it, so we are remarkably near to Pickering & Garrod's "implicit common ground" as proposed in their Interactive Alignment Model (IAM) which in effect states that shared knowledge is by default achieved without iteration of mental states. Somewhat ironically, having started with Clark, we arrive at Pickering & Garrod, the reason being that their interactive alignment model is intention-free in the default case.

and represents an acceptable external datum from the point of view of ethno-methodology and dialogue theory.

Acknowledgements We gratefully acknowledge partial support by the DFG funded SFB 673 “Alignment in Communication” as well as by the German Aerospace Center (DLR) with funds from the Federal Ministry of Economics and Technology (BMBF) due to resolution 50RA1023 of the German Bundestag.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- Barwise J (1989) On the model theory of common knowledge. In: The situation in logic, vol 17, pp 201–221
- Barwise J, Moss L (1996) Vicious circles: on the mathematics of non-wellfounded phenomena. Cambridge University Press, Cambridge. (Center for the Study of Language and Information)
- Bird A, Tobin E (2010) Natural kinds. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (Winter 2012 Edition)
- Clark H (1996) Using language. Cambridge University Press, Cambridge
- Clark H, Brennan S (1991) Grounding in communication. In: Perspectives on socially shared cognition, vol 13, pp 127–149
- Clark H, Marshall C (1981) Definite reference and mutual knowledge. In: Joshi AK, Webber B, Sag I (eds) Elements of Discourse Understanding
- Clark H, Schaefer E (1989) Contributing to discourse. Cogn Sci 13(2):259–294
- Fagin R, Halpern J, Moses Y, Vardi M (1995) Reasoning about knowledge. MIT Press, Cambridge
- Fernandez R, Endriss U (2007) Abstract models for dialogue protocols. J Log Lang Inf 16(2):121–140
- Ginzburg J (2012) The interactive stance. Meaning for conversation. Clarendon Press, Oxford
- Harman G (1977) Review of *linguistic behavior* by Jonathan Bennett. Language 53:417–424
- Lang C, Hanheide M, Lohse M, Wersing H, Sagerer G (2009) Feedback interpretation based on facial expressions in human–robot interaction. In: International symposium on robot and human interactive communication (RO-MAN). IEEE, New York
- Lewis D (1969) Convention. Harvard University Press, Cambridge
- Lütkebohle I, Hegel F, Schulz S, Hackel M, Wrede B, Wachsmuth S, Sagerer G (2010) The Bielefeld anthropomorphic robot head flobi. In: 2010 IEEE international conference on robotics and automation (ICRA). IEEE, New York, pp 3384–3391
- Meyer J, Van Der Hoek W (2004) Epistemic logic for AI and computer science. Cambridge University Press, Cambridge
- Peltason J, Wrede B (2010) Modeling human-robot interaction based on generic interaction patterns. In: AAAI Fall symposium: dialog with robots. AAAI, Menlo Park
- Peltason J, Wrede B (2011) The curious robot as a case-study for comparing dialog systems. AI Mag 32(4):85–99
- Peltason J, Riether N, Wrede B, Lütkebohle I (2012) Talking with robots about objects: a system-level evaluation in HRI. In: International conference on human-robot interaction, HRI'12. ACM/IEEE, New York, pp 479–486
- Traum D (1999) Computational models of grounding in collaborative systems. In: Psychological models of communication in collaborative systems—papers from the AAAI Fall symposium, pp 124–131
- Traum DR (1994) A computational theory of grounding in natural language conversation. PhD thesis, Rochester, NY, USA
- Tuomela R (2000) Cooperation: a philosophical study. Springer, Berlin
- Tuomela R (2002) The philosophy of social practices: a collective acceptance view. Cambridge University Press, Cambridge
- Walker MA, Litman DJ, Kamm CA, Abella A (1997) PARADISE: a framework for evaluating spoken dialogue agents. In: 8th conference of the European Chapter of the Association for Computational Linguistics

Julia Peltason is a research assistant in the Applied Informatics Group at Bielefeld University, Germany, where she is working towards her Ph.D. in the field of human-robot interaction. Her research interests include dialog system engineering, interaction design and evaluation of interactive systems. She has contributed to several projects exploring social robots and robot assistants.

Hannes Rieser is Professor *Emeritus* of Semantics and Pragmatics at the Faculty of Linguistics and Literary Studies, Bielefeld University. His research interest are Dynamic Semantics, Dialogue Theory, gesture research and speech-gesture interfaces. He has published in all these areas.

Sven Wachsmuth is holding a Senior Lecturer position at Bielefeld University and is currently heading the Central Lab Facilities of the Center of Excellence Cognitive Interaction Technology (CITEC). He received the Ph.D. degree in computer science from Bielefeld University in 2001. In 2003, he spent a sabbatical year at the Computer Science Department of the University of Toronto. His research interests are in Human-Robot Interaction, especially looking at high-level computer vision problems, and

system integration and evaluation aspects. He is principal investigator in the collaborative research centre 673 (Alignment in Communication) and has been a member of the technical committees of the RoboCup@Home organization in 2010–2013.

Britta Wrede is currently head of the Applied Informatics Group at Bielefeld University. Her research is driven by the question how to equip robots with a better understanding of their environment and is strongly inspired by human development. It follows the hypothesis that learning needs to be embedded in social interaction.