Spatial Knowledge Representation for Human-Robot Interaction

Reinhard Moratz[1], Thora Tenbrink[2],
John Bateman[3], Kerstin Fischer[3]

[1]University of Bremen, Center for Computer Studies
Bibliothekstr. 1, 28359 Bremen, Germany
moratz@tzi.de

[2]University of Hamburg, Department for Informatics
Vogt-Kölln-Str. 30, 22527 Hamburg
tenbrink@informatik.uni-hamburg.de

[3]University of Bremen, FB10: Linguistics and Literary Studies
Postfach 330440, 28334 Bremen, Germany
{bateman, kerstinf}@uni-bremen.de

Abstract. Non-intuitive styles of interaction between humans and mobile robots still constitute a major barrier to the wider application and acceptance of mobile robot technology. More natural interaction can only be achieved if ways are found of bridging the gap between the forms of spatial knowledge maintained by such robots and the forms of language used by humans to communicate such knowledge. In this paper, we present the beginnings of a computational model for representing spatial knowledge that is appropriate for interaction between humans and mobile robots. Work on spatial reference in human-human communication has established a range of reference systems adopted when referring to objects; we show the extent to which these strategies transfer to the human-robot situation and touch upon the problem of differing perceptual systems. Our results were obtained within an implemented kernel system which permitted the performance of experiments with human test subjects interacting with the system. We show how the results of the experiments can be used to improve the adequacy and the coverage of the system, and highlight necessary directions for future research.

Keywords. Natural human-robot interaction, computational modeling of spatial knowledge, reference systems

1 Introduction and Motivation

Many tasks in the field of service robotics will profit from natural language interfaces that are capable of supporting more ‘natural’ styles of interaction between robot and user. Most typical scenarios include a human user instructing a robot to perform some action on some object. But a precondition for the successful performance of this kind of task is that human and robot establish joint reference to the objects concerned. This requires not only that the scene description created by the robot’s object recognition system and the visual system of the human instructor be matched, but also that the

robot can successfully mediate between the two kinds of description—that is, between its internal spatial representations of the position and identity of objects and the styles of language that humans use for achieving reference to those objects. There are two substantial problem areas facing solutions to this task: one arising out of the very non-human-like perceptual systems employed by current robots, the other out of the fact that human language users rarely employ the complete and unambiguous references to objects that might be naively expected.

In this paper, we present an experimental system that employs a computational model designed for the mapping of human and robotic systems. Based on a more detailed analysis of the results of an exploratory study which has been previously described in [Moratz et al., 2001], we show how the two problem areas at hand need to be addressed, present an expanded version of the earlier computational model that was used in the study, and open up perspectives for necessary future research.

1.1 Two problem areas in achieving spatial reference

The first problem area, concerning the differing perceptual capabilities of humans and robots, gives rise to a range of divergences between strategies found in human-human communication and those applicable to the human-robot domain. Between humans, reference objects can usually be specified by the class name of the object. However, when the robot has no detailed *a priori* knowledge about all of the relevant objects (for example, CAD data, knowledge from a large training set), the current state of the art does not allow correct object categorization by class. Although modern automatic object recognition systems are increasingly good at identifying individual objects if the system has been trained for the specific object features, recognizing known objects is only one important aspect of successful communication between humans and robots. Very often, it is new objects in an open scenario that have to be categorized correctly in order to identify the object referred to by the speaker. This may cause severe communication problems that compromise the interaction. For example, while in human-to-human communication reference is often established by using the object category name, such as: “the key on the floor,” a corresponding natural language human-to-robot instruction that accommodates the perceptual abilities of the robot may need to be more like: “the small reflecting object on the ground, to the left of the brown box.” This is because robots have limited perceptual capabilities that often preclude accurate recognition of broadly similar objects and, moreover, may not have access to the necessary world knowledge that would identify the object by class.

The second problem area, the partiality of human strategies for achieving spatial reference, is clearly shown by work on achieving reference to objects undertaken within the field of natural language generation. Reiter and Dale (1992) , for example, have shown both that the general task of producing a guaranteedly unambiguous referential expression for some object from a set of potential referents is NP-hard and that humans (perhaps as a consequence) do not in any case attempt to construct guaranteedly unambiguous references. The referential strategies adopted in natural human-human communication usually employ perceptual salience, recent mention in the discourse, and deictic modifiers (this, that, etc.) in order to achieve successful reference without requiring a solution to the problem of creating an optimal (i.e., shortest uniquely

identifying) reference expression. Furthermore, since the possibility of error is in-built, interactive strategies are employed by all participants in an interaction in order to provide opportunities both for inobtrusively exhibiting what has been understood and for smoothly correcting misunderstandings that have occurred. Only when reference is still not successful does the human interactant need to resort to explicit correction of the misunderstanding. These interactional techniques have been widely researched, particularly within the conversation analytic tradition [Schegloff et al., 1977].

The relative 'unnaturalness' of the second referring expression used above, which is the perceptually appropriate one for the robot, is then a direct consequence of these properties: for a natural interaction the expression sounds both over-explicit ("small reflecting ... on the ground") and under-specific ("object" instead of "key"). Ways need to be found of ameliorating both problems if more natural interactive styles are to be achieved.

1.2 Qualitative spatial reference as a communicative device

To address these problems, a different, powerful strategy for achieving reference in human-human communication can be considered more closely. Whereas many objects may have some particular color, size or texture—which therefore give rise to more potential confusion, or 'distractors', for a referential expression—the position of objects is generally uniquely defining; if identified sufficiently restrictively, only one object is in a given place at a time. This could make the use of explicit positional information a good strategy for achieving unique reference in the human-robot communicative situation also. However, specifying positional information in the human-robot context also faces the above problems of mismatched perceptual systems (the robot is good at exact range-finding, humans are not) and object identification (relative positions need to reference other objects, and these objects again must be identifiable by the robot). Therefore, although more constrained, the problems above still need addressing. In this paper, we focus particularly on this use of positional information for reference to objects in human-robot interaction and attempt appropriate solutions to the accompanying problems by means of empirically deriving a spatial representation supportive of natural interaction. Qualitative spatial reference then serves as a necessary bridge between the metric knowledge required by the robot, and more 'vague' concepts that build the basis for natural linguistic utterances, as suggested by Hernández (1994) . As the kinds of spatial representations and appropriate language forms to be adopted still need to be ascertained and evaluated empirically, we investigate this area further in an exploratory study. The results are outlined in detail in section 4.

In our scenario, a human user is asked to instruct a robot to move to one of several similar objects that are arranged in the spatial vicinity of the robot and in some cases also in the vicinity of a further, different object. The robot is equipped with a prototypical object recognition system characterized by the following features: The system can determine the metrical position (distances and angles) relative to the robot, and estimate the approximate size of an object in relation to the robot's size—i.e., larger or smaller than the robot; it can make a coarse classification of the object's shape (compact vs. long); and it can provide coarse colour information (ca. six to eight colour categories,

although sharp distinctions between categories such as **red** and **orange** are not available). However, the system is unable to deal with gestural indications of direction. Furthermore, due to these system limitations the robot is not able to make fine distinctions between roughly similar objects. Thus, it will not be able to identify objects correctly on the basis of a human instructor's verbal input if such input refers to fine-grained non-positional differences between the objects in question. The simple experimental configuration then forces the human user to explore other ways of referring to objects and, here, distinguishing objects on the basis of their position in space becomes a natural candidate. However, as humans in natural surroundings are not capable of providing exact metrical information about distances and angles, the objects' positions have to be referred to by qualitative information such as their relative position and other referential strategies. Ascertaining these strategies and their effectiveness was then one goal of our experimental set-up.

To formulate hypotheses about the expected user strategies in qualitative linguistic spatial reference, we can draw on previous research (e.g., [Levinson, 1996]) on human strategies for achieving such reference within naturally occurring scenarios to a certain degree. The perspective used in our scenario is, however, fundamentally different to that in most human-human interaction scenarios. In a typical experiment carried out to trigger human subjects' linguistic references, a relevant question could be: "Where is the object?". A typical answer describes the object's location by referring to its spatial relation to other available entities, such as the speaker, the hearer, or another object. In contrast, the restrictions we have seen in human-robot interaction readily create the need to refer to objects in ways which are less common in natural human-human interaction. Using the positional strategy for reference, for example, reverses this last perspective: it is not the position of an object that is unknown, but rather the *identity* of one of several entities with known positions. Thus, the issue at hand becomes: "Which of these similar objects are you referring to?". This scenario triggers strategies of linguistic reference hitherto largely ignored in the literature on spatial reference systems. We have accordingly adopted a very constrained scenario that effectively forces interaction of the kind required.

1.3 Application in a situated and integrated instruction scenario

A spatial and instructional knowledge representation serves as point of integration for the robot's language and vision mediated information. This provides for an integrative and coherent representation of objects, events and facts. Both modalities, language and vision, are then made available to the processes of understanding via a common representation level.

Central to the architecture of our experimental system is the insight that spatial instruction is:

> **situated**: the discourse relates to a scene which can be understood using only limited previous knowledge. The visual access to a mutually perceived scene supports a state of joint attention to real-world objects.
>
> **integrated**: the integration of language and vision as well as action allows a single consistent interpretation.

This architecture will be discussed below. These two central factors, situatedness and integratedness, determine the procedure used in this paper.

1.4 Related research

Natural language is now established as a crucial component of any 'natural', user-friendly and appropriate interface supporting communication between computational systems and their human users. The integration of language and perception has a long tradition ([Neumann and Novak, 1983], [Wahlster et al., 1983], [Hildebrandt et al., 1995], [Moratz et al., 1995]). In the context of spatial robot-human interaction, natural language performs several particularly important interactional functions: e.g., task specification, monitoring, explanation/recovery, environment update/refinement (e.g. Stopp et al. (1994)). While mostly not focusing on spatial aspects of the interaction, many current research efforts are attempting to improve the naturalness and ease of such communication; projects such as Morpha (BMBF), SFB 360 (Bielefeld), SFB 378 (Saarbrücken) all give dialog a central place and consider it a necessary feature of robot-human interaction. Each project places different priorities and emphases on different aspects of dialog. The situation is very similar for assistance systems, such as SmartKom. Also within these projects, however, the linguistic channel is combined with interaction via graphics, gestures and the like (e.g., Streit (2001) , Lay et al. (2001) , Wahlster (2001)). This is undoubtedly important and will significantly shape the interfaces of the future. However, there are situations where the augmentation and replacement of natural language based interaction through graphical, gestural and other channels is not possible or appropriate. In such cases, as in our scenario, the only access to the robotic system the user has is the linguistic channel.

While these research areas provide valuable contributions to the questions addressed in this paper, the specific effect of a *robot* interaction partner on the linguistic and spatial choices of a human speaker has not been addressed so far. As previous studies in the related field of human-computer interaction, e.g. [Amalberti et al., 1993], [Fischer, 2000], have shown, such specific effects are highly probable, as the users' conceptualisation of their interaction partners has considerable impact on their language. Moreover, due to the situatedness and integratedness of the communication situation the user is focused on the interaction situation itself, which increases the influence of specific situational variables. The question of which spatial reference systems are employed by speakers under which circumstances when interacting with a robot therefore still needs further exploration.

In the next section, we sketch the variability of qualitative spatial reference systems available to humans when referring to a (visually available) object's position. Then, we describe the natural language controlled robot system that we used in our exploratory study in human-robot interaction, which is presented in section 4. The data elicited allow determining the range of spatial instructions used, showing the situatedness and integratedness of the instructions. For the experiments, a preliminary version of our computational model was used, which is described in detail in [Moratz et al., 2001]. In section 5, we present a redesign of the model, which is based on the findings of our experiment and which accounts for the range of representational choices employed by

the users. Finally, we open up perspectives for future work for using human spatial reference in the interaction with robotic systems.

2 Spatial Instructions Using Intrinsic, Relative, and Absolute Reference Systems

Previous research on reference systems employed by humans for locating one object in relation to another object of a different natural kind (cf. [Levinson, 1996] and [Herrmann, 1990]) has led to the identification of three different reference systems, termed by Levinson (1996) *intrinsic*, *relative*, and *absolute*. Each of these occurs in three further variations dependent on whether the speaker, the hearer, or a third entity serves as the origin of the perspective employed. In this section, we start from this classification of spatial reference systems in order to apply it to our specific scenario involving the identification of one of several similar objects rather than the localisation of one object. Here, objects may be classified (i.e. perceptually grouped) into and referred to as groups rather than individual objects. The position of one of the objects may then be referred to by determining its position relative to the rest of the group. Such a scenario is rather typical in human-robot interaction, but has been largely ignored in previous research on linguistic spatial reference. We offer an expansion of well-established classifications of spatial reference systems to address the question how one member of a group of objects is identified[1]. Furthermore, we use several applicable results from previous psycholinguistic research to formulate assumptions about which options of the variety of reference systems theoretically available to speakers can be expected to be employed by the users in our scenario.

In *intrinsic reference systems*, the relative position of one object (the *referent*) to another (the *relatum*) is described by referring to the relatum's intrinsic properties such as *front* or *back*. Thus, in a scenario where a stone (the referent) is situated in front of a house (the relatum), the stone can be unambiguously identified by referring to the house's front as the origin of the reference system: "The stone is in front of the house". In such a situation, the speaker's or hearer's position are irrelevant for the identification of the object. However, the speaker's or hearer's front or back, or, for that matter, left or right, may also serve as origins in intrinsic reference systems: "The stone is in front of you". In such cases, no further entity (such as, in our example, the house) is needed, which is why Herrmann (1990) refers to this option as *two-point localisation*.

In a scenario where *groups of objects* serve as relatum, they can only be used for an intrinsic reference system if they have an intrinsic front. For example, to identify one person in a group of people walking in one direction one could refer to "the one who walks in the front of the group."

Humans employing *relative reference systems*, or, in Herrmann's terminology, *three-point localisation*, use the position of a third entity as origin instead of referring to in-

[1] Apart from the need for expansion of previous accounts, it is necessary to be very explicit about the terminology employed in our approach, as the literature on spatial reference systems is full of ill-defined, overlapping, or conflicting usages of terms. For instance, we avoid the term *deictic* as used, among others, by [Retz-Schmidt, 1988], as it has been variously used to denote contradicting concepts, see [Levinson, 1996].

built features of the relatum. Thus, the stone (the referent) may be situated to the left of the house (the relatum) from the speaker's, the hearer's, or a further entity's point of view (origin): "Viewed from the hut, the stone is to the left of the house". Here, the house's front and back are irrelevant, which is why this reference system can be employed whenever the position of an object needs to be specified relative to an entity (a relatum) with no intrinsic directions, such as a box.

If the stone is related to *a group of* other stones, it may be situated, for instance, to the left of the rest of the group, and this may be true from the speaker's, the hearer's, or a third entity's point of view. A typical example would be, "the leftmost stone from your point of view".

In *absolute reference systems*, neither a third entity nor intrinsic features are used for reference. Instead, the earth's cardinal directions such as *north* and *south* (or, in some languages, properties such as *uphill* or *downhill* [Levinson, 1996]) serve as anchor directions. Thus, the stone may be to the north of the speaker, the hearer, or the house. Equivalently, if the stone is situated in a *group of stones*, it may be located to the north of the rest of the group. Absolute reference systems are a special case in that there is no way of labelling "origins" or "relata" in a way consistent with the other kinds of reference systems, as directions behave differently than entities.

For our experimental scenario the following initial assumptions can be made. Although humans generally use their own point of view in spatial reference, they usually adopt their interlocutor's perspective if action by the listener or different cognitive abilities on the part of the listener are involved [Herrmann and Grabowski, 1994] . Both of these factors are true in our scenario; therefore, speakers are likely to use the robot's perspective in their instructions. Furthermore, speakers will disprefer absolute reference systems as these are rarely used in natural human-human interaction in Western culture in indoor scenarios (as opposed, for instance, to Tzeltal [Levinson, 1996], [Levelt, 1996]) [2].

Accordingly, out of the various kinds and combinations of reference systems described above, only three kinds of linguistic spatial reference are likely to be used for communication in our scenario: First, the speakers may employ an intrinsic reference system using the robot's position as both relatum and origin. In this case, they specify the object's position relative to the robot's front. Secondly, they can refer to a salient object, if available, as relatum in a relative reference system, in which case they specify the object's position relative to the salient object from the robot's point of view. Finally, they may refer to the group as relatum in a relative reference system. In this case, they specify the object's position relative to the rest of the group from the robot's point of view.

3 The Natural Language Controlled Robot System

The architecture of the system used for experimentation is described in detail in [Habel et al., 1999] . We summarize here the main properties of the system's components.

[2] While, to our knowledge, this intuition has not been directly addressed experimentally, it can be derived from the literature on the kinds of spatial reference systems used by humans in diverse scenarios.

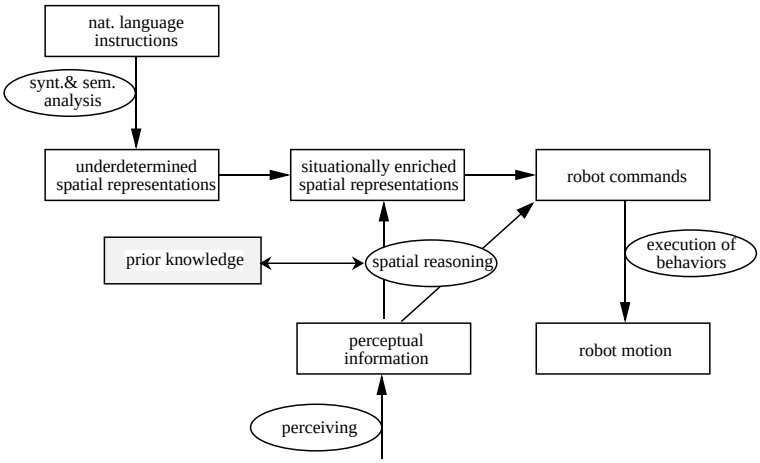

Fig. 1. Coarse architecture of a NL-instructable robot: modules and representations (from [Habel et al., 1999])

The following components interact: the syntactic component, the semantic component, the spatial reasoning component, and the sensing and action component (see figure 1). We can see from the architecture a relatively traditional view of the role of language in robot control in that it is assumed that the human user gives sufficiently clear and unambiguous instructions for the robot to act upon; as we have suggested, for complex reference tasks this is unlikely unless the user is specifically requested to perform in this way (and even then they might not be very good at it). This simplification is appropriate for our experimental purposes, however, in that it forces the user to work through the range of referential strategies naturally available (see section 4).

The *syntactic component* is based on Combinatory Categorial Grammar (CCG), developed by Steedman and others (cf.[Steedman, 1996]). The syntactic component was developed as part of SFB 360 at the University of Bielefeld [Moratz and Hildebrandt, 1998] , [Hildebrandt and Eikmeyer, 1999]. The output of the syntactic component consists of feature-value structures.

On the basis of these feature-value structures, the *semantic component* produces underspecified propositional representations of the spatial domain. In the exploratory study, this component uses a first version of our computational model of projective relations, which is described in more detail in [Moratz et al., 2001]. In section 5, we present an extended version of this model which is based on the results gained in the study. The model maps the spatial reference expressions of the given command to the relational description delivered from the sensor component.

The *spatial reasoning component* plans routes through the physical environment. To follow an instruction, the goal representation constructed by the semantic component is mapped onto the perceived spatial context.

The *sensing and action component* consists of two subcomponents: visual perception and behavior execution. The *visual perception subcomponent* uses a video camera. An important decision was to orient to cognitive adequacy in the design of the communicative behavior of the robot, using sensory equipment that resembles human sensorial

Fig. 2. Our Robot GIRAFFE.

capabilities [Moratz, 1997]. Therefore the camera is fixed on top of a pole with a wide angle lens looking below to the close area in front of the robot (see figure 2). The images are processed with region-based object recognition [Moratz, 1997]. The spatial arrangement of these regions is delivered to the spatial reasoning component as a qualitative relational description. The *behavior execution subcomponent* manages the control of the mobile robot (Pioneer 1). This subcomponent leads the robot to perform turns and straight movements as its basic motoric actions. These actions are carried out as the result of passing a control sequence to the motors.

The interaction between the components consists in a superior instruction-reaction cycle between both language components and the spatial reasoning component. Subordinate to this cycle is a perception-action cycle started by the spatial reasoning component, which assumes the planning function and which controls the sensing and action component.

An example from our application illustrates the interaction of the components and the central role of the spatial representation as follows. The command “fahre zum linken Ball” (“drive to the lefthand ball”)[3] is semantically interpreted as shown in figure 3.

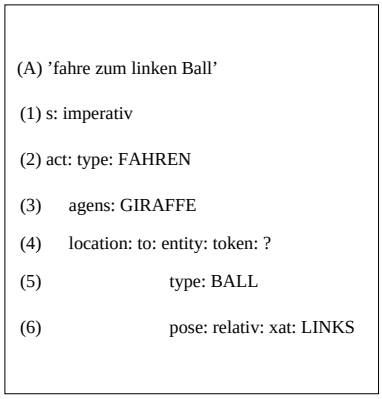

Fig. 3. Semantic interpretation

Now an object that denotes “the lefthand ball” has to be found in the perceived scene. There is a configuration of two balls one of which is to the left of the centroid of the group seen from the robot. This ball is identified as the goal of the robot. Since there is no obstacle, the action invoked will be a direct goal approach to execute the users’ command.

More complex path planning is necessary for finding paths around obstacles. To achieve this, the visual perception subcomponent has to localise the objects, and the spatial reasoning component needs to find some suitable space for movement in order to establish a qualitative route graph.

4 Exploratory study

Our exploratory study was carried out for three primary reasons:

- Human users do not necessarily employ spatial instructions that robots can understand, and they may use strategies for spatial instruction that are different from those investigated in human-to-human communication. One aim was therefore to collect instances of spatial instructions actually employed by users in a human-robot interaction scenario.
- Since spatial instruction is situated, integrated, and involves (at least) two discourse participants, humans approach spatial instruction in an interactive way, using the situation, the actions involved as well as the kinds of sensory input available, and

[3] Translations are approximations and have to be treated with caution. In the mapping of spatial reference systems to linguistic expressions, there is no one-to-one correspondence between English and German.

the possibility of interaction as a resource for their verbal instructions. We therefore aimed at working out ways in which human-robot communication is situated, integrated, and interactive.
– A third aim was to test the adequacy of the implemented version of our computational model with regard to the kinds of spatial reference systems employed by the users.

In the following section, the experimental set-up is described, section 4.2 then describes the results. Subsequent sections describe the primary uses then made of the experimental results.

4.1 Setting

The exploratory study involved a scenario in which humans were asked to instruct our Pioneer 1 robot GIRAFFE (**G**eometric **I**nference **R**obot **A**dequate **F**or **F**loor **E**xploration, see figure 2) to move to one of several roughly similar objects. The experimenter used only pointing gestures to show the users which goal object the robot should move to; pointing was used in order to avoid verbal expressions or pictures of the scene that could impose a particular perspective, for example, a view from above. Users were instructed to use natural language sentences typed into a computer to move the robot; they were seated in front of a computer in which they typed their instructions. The users' perception of the scene was one in which a number of cubes were placed on the floor together with the robot, which was set up at a 90 degree angle or opposite to the user, as shown in figure 4. The fixed setting allows the analysis of the point of view taken by the participant depending on the instructions used. The arrangement of the cubes was varied, and in some of the settings, a cardboard box was added to the setting in order to trigger instructions referring to the box as a salient object.

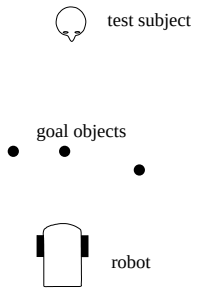

Fig. 4. The experimental setting

As outlined above, the robot can understand qualitative linguistic instructions, such as "go to the block on the right". If a command was successful, the robot moved to the block it had identified. The only other possible response was "error". This disabling of the natural interactive strategies of reference identification challenged users to try out many different kinds of spatial instruction to enable the robot to identify the intended aim. We were therefore able to obtain both a relatively complete indication of the kinds

of strategies available to human users with respect to this task and an indication of the user's willingness to adopt them. 15 different participants carried out an average of 30 attempts to move the robot within about 30 minutes time each. Altogether 476 instructions were elicited.

4.2 Experimental Results

Throughout the experiments, the participants employed the robot's perspective, i.e., there were virtually no instructions in which the user expected the robot to use a reference system based on the speaker or a further object as origin (except for one case in which after a mistake the user explicitly stated that she assumed the robot to be using her point of view). Furthermore, whenever the users referred to the goal object, they overwhelmingly used basic level object names such as *Würfel* [**cube**], and there was also a very consistent usage of imperatives rather than other, more polite, verb forms.

However, the participants in the experiment nevertheless showed considerable variation with regard to the instructional strategies employed. Half of the participants started by referring directly to the goal object, using instructions such as *fahr bis zum rechten Würfel* [**drive up to the right cube**]. When instructions of this kind were not successful—because of orthographic, lexical, or syntactic problems—the participants turned to directional instructions; if successful, they re-used this goal-naming strategy in later instructions. The other half of the participants started by describing the direction the robot had to take, for instance, *fahr 1 Meter geradeaus* [**drive 1 meter straight ahead**]. If they were unsuccessful with this type of instruction, some users turned to decomposing the action into even more detailed levels of granularity, using instructions such as *Dreh dein rechtes Rad* [**turn your right wheel**].

This pattern of usage reveals an implicational hierarchy among the adopted strategies. On reaching a failure, users would change their strategy only in the direction of expected 'simplicity'; they would not attempt a strategy with expected 'higher' complexity. Thus, a fixed order of instructional strategies became apparent which can be roughly characterized as Goal - Direction - Minor actions. This is an important result for designing human-robot interaction—not least because the notion of 'simplicity' maintained by a user need not relate at all to what is actually simpler for a robot to comprehend and carry out. Thus attempts on the part of the user to provide 'simpler' instructions may in fact turn out to confuse rather than aid the situation.[4] Such mismatches can therefore lead to insoluble dialogic problems that are particularly frustrating for users, since they believe (mistakenly) that they are making things easier for the robot. Thus, in the future dialogue components will need to be designed that can detect such a situation and then correct the user's underlying assumptions unobtrusively.

In the following, we analyse in detail the kinds of spatial reference systems employed in these different kinds of instruction. As our aim was to explore the range of instructions employed by the users, and to analyse their instructional strategies on a qualitative level, we did not attempt to work out user preferences quantitatively, using

[4] A related instance of this problem has also been noted when attempting to have users produce more intelligible speech. This can easily lead a user to 'hyper-articulate' which reduces the reliability of speech recognition still further [Oviatt et al., 1998].

statistical measures. However, for illustration of the tendencies we worked out, we add the absolute numbers of occurrence.

Goal instructions (183 occurrences) Spatial instructions indicating the position of the goal object are identified as bounded linear oriented structures in [Eschenbach et al., 2000] . They include directional prepositional phrases specifying the end of the path. Out of the 183 linguistic instructions collected in our experiment that refer directly to the goal object, 102 utterances use the group as a whole as relatum, identifying the intended object by its position relative to the other objects in the group. 69 of these 102 group-based references used a particular expression schema consisting of an imperative combined with a locative directional adjunct specifying relative position; as, for example, in *Fahr zum linken Würfel* [**Drive to the lefthand cube**] where the locative adjunct gives the relative position of the cube in the group to which it belongs. The lexical slots for the verb and object in this schema were varied, as were the positional adjective of the locative adjunct—yielding "mittleren" [**middle**], "hinteren" [**back**], "vorderen" [**front**] in addition to "linken" [**left**].

For some situations, besides the cubes used as goal objects the setting included a further object, namely a cardboard box which could be used as a reference object. In 19 cases of the 43 instructions uttered in situations where this salient object was present, the cardboard box was used for a relative reference system with the salient object as relatum. Here, the syntactic structure used most often is also quite stable: an imperative and two hypotactic adjuncts are used, with the subordinated adjunct identifying the relatum's position relative to the adjunct specifying the reference object, as in: *geh zum Würfel rechts des Kartons* [**go to the cube to the right of the box**].

The robot's intrinsic properties are used for instruction in altogether 42 of the 183 goal-oriented instructions, using various linguistic expressions such as *Fahr zum Würfel rechts von dir* [**Drive to the cube to your right**]. Although the orientation of the robot is not stated explicitly in these commands, the speakers could not use an expression like "to your right" without assuming a front of the robot.

Altogether, these results correspond to the expectations we outlined in section 2. Those users who referred to the goal object all employed the three kinds of reference systems expected, and they consistently used the robot's perspective (which is actually a more homogeneous usage than we might expect). Strikingly, in all of the goal instructions except for those employing the robot's intrinsic properties, the users failed to specify the point of view they employed, rendering the instructions formally ambiguous with regard to the variability of origins but, we would claim, appropriate within the particular situated interaction.

Direction instructions (210 occurrences) In altogether 210 instructions, the goal object is not specified directly, but a direction of movement is indicated. In more than half of these instructions, a verb of locomotion such as *fahre* [**drive**] or *rolle* [**roll**] is used; the others simply specify the direction itself. This variability does not reflect any relevant semantic differences (the only way the robot can move at all is by using its wheels) and are therefore not discussed further here. Other verbs of motion such as verbs of transport ("bring"), change of position ("enter") and caused change of position

("put") (cf. [Eschenbach et al., 2000]) do not occur in this simple scenario. Directional instructions indicate unbounded linear oriented structures, as only the initial step of an intended goal-directed path is expressed. No further steps occur in this scenario for lack of reaction by the robot. As an exception, two instructions may be combined in one utterance (see below), but these still do not include a goal, i.e., the structures are still unbounded. In more than half of the directional instructions (141 out of 210), the intrinsic point of view of the robot is used as origin of a reference system which employs the principal directions as defined in [Eschenbach, 2001]. In 78 of these cases, these principal directions are employed without modifications, as in: *vorwärts* [**forward**], and *gehe nach links* [**go to the left**]. "Vorwärts" expresses the standard orientation of a body during motion, i.e., the alignment of the object order of the path with the intrinsic front-back axis of the robot (cf. [Eschenbach, 2001]). Several users employed the earth's cardinal directions (12 occurrences) rather than relying on the principal directions based on the robot's physical properties, as in *Gehe nach Norden* [**Go to the North**]. Altogether, in almost half (90 out of 210) of the directional, non-goal specifying instructions, the users indicated an unmodified principal or absolute direction to make the robot move, obviously leaving further specifications of the path for later instructions.

Nevertheless, many users seemed to assume that the intended goal was not directly accessible by simply moving in one of these cardinal directions. Thus, in 32 instructions the angle in which the robot should move is specified more exactly, using either quantitative (8 occurrences) measures such as *20 Grad nach rechts* [**20 degrees to the right**] or qualitative (24 occurrences) specifications, for instance, *geradeaus etwas rechts fahren* [**drive forward somewhat to the right**]. One-third of these instructions employed a combination of either a principal direction and an angle (in quantitative usages), or two principal directions (in qualitative usages). Some users explicitly divided such a combination into two partial instructions (4 occurrences) which were to be carried out one after the other, as in *gehe vorwärts dann nach rechts* [**move forward then to the right**].

Some users indicated the length of the intended path, using either quantitative (18 occurrences) measures such as *Fahre 1 meter geradeaus* [**Drive forward 1 meter**], or qualitative (8 occurrences) expressions such as *Fahre ein wenig nach vorn* [**Drive a bit forward**]. Interestingly, in contrast to the findings on angle specifications, in this case the quantitative instructions outweighed the qualitative ones. One user tried out an instruction specifying not only the direction but also the length of time during which the robot was supposed to move in that direction: *Fahre 1 Sekunde vorwärts* [**Drive forward 1 second**]. Some of the instructions (52 occurrences) relied on a different, salient entity (a landmark) available in the room for specifying the intended path rather than relying on the principal directions determined by the robot's intrinsic properties. Of these 52 instructions, 46 referred to the cardboard box which was available only in some of the scenarios, as in: *umfahre den Kasten* [**drive around the box**]. Mostly, these instructions (in contrast to the goal-based instructions) do not command the robot to move *to* the box, but rather around it, behind it, or beside it. Thus, it is linguistically expressed that the box is not itself the intended goal. The other 6 instructions used

entities located at a greater distance from the robot to specify the intended direction, as in *Fahre zur Wand* [**Drive to the wall**].

Finally, in a few (4) instructions the users left it to the robot to decide about the correct orientation, as in *Fahre im Kreis* [**Drive in a circle**].

Minor action instructions (83 occurrences) The remaining 83 instructions did not specify either the goal object or a direction in which the robot should move, but instead decomposed the action into minor activities. In 28 of these instructions, the users did not command the robot to **move** in a direction, but rather to change its orientation into a specific direction, as in *dreh dich nach rechts* [**turn to the right**].[5] About half of these instructions involved qualitative, the other half quantitative measures. 29 instructions indicated that the robot should move, but were confined to the verbs of locomotion, such as *Fahren* [**Drive**]. The remaining 26 instructions reflected the users' individual, sometimes rather desperate attempts to communicate with the robot at all, as exemplified by utterances such as *Tu was* [**Do something**] and *Schalte den Motor ein* [**Turn on the engine**].

5 A revised computational model for the spatial human-robot interaction scenario

The experiments described above, which were carried out using a previous version of the system, provided several valuable clues as to how a new system could be designed. A new system is currently being set up, and instead of using the GIRAFFE platform, this new system will use the Sony AIBO robotic system (see figure 5). Because of its animal-like shape, the AIBO might be perceived as a more natural communication partner. In addition, the AIBO is also well suited for gathering data on robots used in an entertainment context.

However, the four-legged robot AIBO and the GIRAFFE differ regarding their respective perceptual abilities (field vs. survey perspective; orientation knowledge vs. position - i.e., orientation and distance - knowledge), which might trigger various kinds of interesting communication problems. Because the AIBO camera in its head is closer to the ground than that of GIRAFFE, the AIBO is unable to calculate precise distances to unrecognized objects. Thus AIBO has only orientation knowledge available to it, and this knowledge has to tally with the survey knowledge provided by the human instructor. By changing its position, the AIBO acquires new perspectives and further orientation information on the scene. A spatial inference engine combines the information from the AIBO's different viewpoints, along with the survey knowledge provided verbally by the human instructor, to build up a depiction of the environment. In order to draw spatial inferences using the spatial inference engine, the verbal description provided by the human instructor must first be transferred into a spatial reasoning calculus (QSR calculus). For our system, we employ the TPCC calculus, introduced in the current volume (see Moratz, Nebel, Freksa (2002)).

[5] These are not counted as directional movement instructions as they express an action on a finer level of granularity, leaving out locomotion.

Fig. 5. Legged AIBO robot.

Given the results of our experiments, and building on the general results from psychology and psycholinguistics on spatial expressions in human-to-human communication that we summarized in section 2 above, it was possible to design a level of representation that provides our robot with a model of the verbal strategies of spatial instructions produced by users in the experimental scenario. This model consists of two parts: first, a knowledge base representing the coarse structure and links to general world knowledge (section 5.1); second, a representation capturing the fine grained positional information (section 5.2) represented using the TPCC calculus [Moratz et al., 2002]. The knowledge base offers a blueprint from which individual spatial instructions can be derived as particular instances. Such instances then provide the necessary link between the language input module and the navigation module presented in section 3 above.

5.1 The semantic structure of spatial instructions for mobile robots

The representation formalism we adopt is derived from the ERNEST system ([Niemann et al., 1990] , [Kummert et al., 1993], [Moratz, 1997]). ERNEST is a semantic network formalism in the KL-ONE tradition, providing a subset of representation and inference capabilities relevant for robotic reasoning. It can be used for the representation of concepts and the relationships between them and has already been applied successfully in the context of integration of linguistic and perceptive knowledge [Hildebrandt et al., 1995]. Since we do not use the inference mechanisms but only the declarative component we can work with a simplified version of ERNEST, which we present here in a short sketch.

The primary elements of an ERNEST semantic network are concepts, their attributes and the relations between concepts. These are usually represented as nodes,

their internal structures and links between nodes respectively. We use two types of nodes:

- A *concept* can represent a class of objects, events or abstract conceptions.
- An *instance* is understood as the concrete realisation of a concept in the input data; i.e. an instance is the copy of a concept by which the general description is replaced by concrete values.

Subordinate features of a concept, such as the size of an object or its colour, are represented by means of attributes. Concepts are therefore entities with internal structure. Features of concepts that are important for the domain are represented as links to other concepts. ERNEST supports the following standard link types:

- Through the link type *role*, two concepts are connected with each other if one concept is understood as a prerequisite of the other.
- Through the link type *specialisation* and a related inheritance mechanism, a special concept is stated to inherit all attributes and roles of the general one.

The knowledge present in the semantic network is utilized by creating instances. This process requires that a complex object be recognized as an instance of a concept, which in turn requires that all its necessary roles can be recognized.

The experimental results indicate that certain kinds of information concerning spatial directions commonly occur together and others less so. This was modeled as the semantic network fragment shown in figure 6. In the figure, specialisation links are oriented horizontally and role links are oriented vertically. Optional role links (i.e., the cardinality range includes zero) are shown dashed in the figure. The three main types of instructions found empirically constitute the three specializations shown for the concept 'drive instruction'; the presence of a goal-object (as a subconcept of spatial-object) and of a landmark (a further subconcept of spatial-object) in their respective instruction types are shown by the vertical role links. The obligatory relationship expressed between 'relative position' and 'orientation' is the link to the 'projective-expression' concept which is the interface to the model of projective relations presented in the next subsection.

Instances formed from these concepts interface directly with the robot's control components. Thus, the recognition of a linguistic instruction is responded to by a corresponding action on the part of the robot. Particular instances also have information added via the robot's perceptive apparatus; for example, exact position relative to the robot and basic attributes of colour and size as mentioned above. We will return to some further possible uses of this additional information in section 6 below.

5.2 The interpretation of projective relations

An essential aspect of the robot's ability to execute instructions is its interpretation of the spatial relations specified between objects functioning as landmarks or relatum and the goal objects. The experimental results have a number of consequences for our model of the projective relations and their uses. The computational model shown in figure 6 represents the different kinds of reference systems required for interpreting linguistic

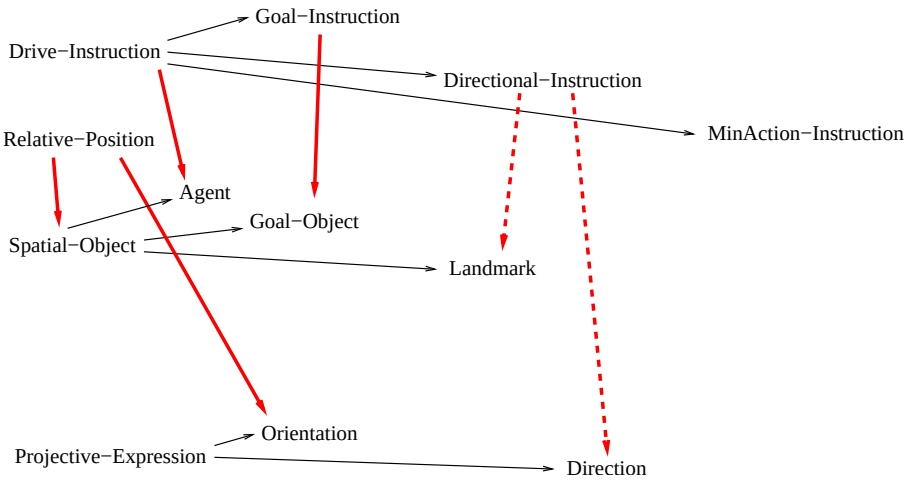

Fig. 6. Knowledge base for the semantic structure of spatial instructions.

references according to the three options outlined in section 2 and for handling the corresponding instructions. Note that our empirical results already allow us to exclude several theoretically possible alternatives that were not, in fact, selected as strategies by our experimental participants: for example, intrinsic and relative reference systems employing either the speaker or a salient object as origin.

The projective expressions are then further resolved as follows. To model reference systems that take the robot's point of view as origin, all objects are represented in an arrangement resembling a plan view (a scene from above). This amounts to a projection of the objects onto the plane $\mathcal{D}$ on which the robot can move. The projection of an object O onto the plane $\mathcal{D}$ is called $p_{\mathcal{D}}(O)$. The center μ of the projected area can be used as point-like representation O' of the object O: $O' = \mu(p_{\mathcal{D}}(O))$. The reference axis is then a directed line through the center of the object used as relatum (see figure 7), which may be the robot itself, the group of objects, or other salient objects.

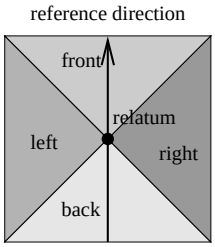

Fig. 7. Relatum and reference direction

The partitioning into sectors of equal size is a sensible model for the directions “links” (left), “rechts” (right), “vor” (front) and hinter (back) relative to the relatum. However, this representation only applies if the robot serves as both relatum and origin. If a salient object or the group is employed as the relatum, front and back are exchanged, relative to the reference direction [Herrmann, 1990]. The result is a qualitative distinction, as suggested, for instance, by Hernandez (1994) . An example for this configuration is shown in figure 8. In this variant of relative localisation, the “in front of” sector is directed towards the robot.

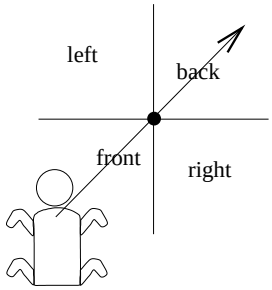

Fig. 8. Relative reference model

In cases with a group of similar objects, the centroid of the group serves as virtual relatum. Here the reference direction is given by the directed straight line from the robot center to the group centroid. The object closest to the group centroid can be referred to as the “middle object” (see figure 9).

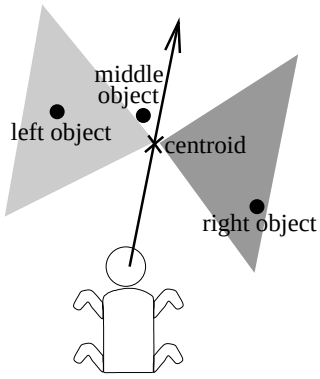

Fig. 9. Group based references

For combined expressions like “links vor” (left in front of) vs. precise expressions like “genau vor” (straight in front of) we use the partition presented in figure 10. This

partitioning can account for the projective expressions used for the orientation in goal instructions as well as the directions in directional instructions (see figure 6 above), in which the robot's position and physical orientation provide the basis for determining the intended reference direction.

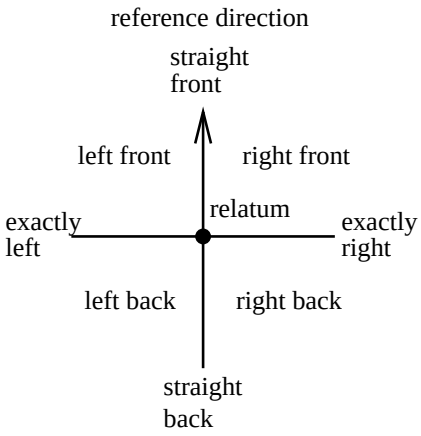

Fig. 10. Model for combined expressions

To define the partitions formally, we refer to the angle ϕ between the reference direction and the straight line from the relatum to the referent, or, respectively, the denoted direction.

$$
\begin{aligned}
referent \text{ vor } relatum &:= -\pi/4 \leq \phi \leq \pi/4 \\
referent \text{ links } relatum &:= \pi/4 < \phi < 3/4\pi \\
referent \text{ hinter } relatum &:= 3/4\pi \leq \phi \leq 5/4\pi \\
referent \text{ rechts } relatum &:= -\pi/4 > \phi > -3/4\pi \\
referent \text{ links} - \text{vor } relatum &:= 0 < \phi < \pi/2 \\
referent \text{ links} - \text{hinter } relatum &:= \pi/2 < \phi < \pi \\
referent \text{ rechts} - \text{vor } relatum &:= 0 > \phi > -\pi/2 \\
referent \text{ rechts} - \text{hinter } relatum &:= -\pi/2 > \phi > -\pi \\
referent \text{ genau} - \text{vor } relatum &:= \phi = 0 \\
referent \text{ exakt} - \text{links } relatum &:= \phi = \pi/2 \\
referent \text{ genau} - \text{hinter } relatum &:= \phi = \pi \\
referent \text{ exakt} - \text{rechts } relatum &:= \phi = -\pi/2
\end{aligned}
$$

The partitions described above exactly correspond to the acceptance areas used in the QSR calculus TPCC (this volume [Moratz et al., 2002]). With the aid of these acceptance areas, the instructor's verbal spatial description information can be matched

to the perceptually captured local view information from the AIBO. One difficulty inherent in this process is that the local view information captured by the AIBO contains only orientation knowledge, lacking distance information. However, the knowledge represented in TPCC can be combined using constraint propagation, and thus it is possible to generate survey knowledge from local knowledge.

6 First steps towards more natural interaction

The design and implementation of the mobile robot GIRAFFE reported so far has already achieved the integration of several different informational modalities. Linguistic input, perception and robot action all combine in the robot's interpretation and execution of the instructions it receives. The implemented model performs adequately in that its primary behavioral mode, following goal-centered instructions, corresponds to the instruction strategy most preferred by users. Users overwhelmingly employed the robot's perspective and most of the spatial reference systems employed corresponded directly to those implemented so that successful communication was achieved. Moreover, based on the fact that there were situations in which other strategies were employed, such as directional instructions or specifications of minor actions, our inclusion of these in the model will allow successful interaction here also.

The experimental results of course raise many more issues. In particular, we consider in this section a more sophisticated use of spatial representation in order to allow successful operation in more demanding circumstances. The correct achievement of reference in human-human interaction is often more negotiated and interactively mediated than was supported by our experimental scenario. However, given the representation uncovered, we are now exploring ways of using interaction to clarify underlying misconceptions on the part of the user such as that which led some of our test persons to believe that they could not directly refer to the intended goal object; and to allow more robust and powerful recognition by the robot of a user's instructions. This can be clarified with some simple examples. Goal objects can currently be recognized on the basis of the linguistic input to the system only when there are not too many competing potential referents. If, for instance, there are several cubes 'to the left' within a group of cubes, then simple reference may fail. Moreover, the very fact that there is a more complex situation for which a user must construct an appropriate referring expression can lead to the production of language that falls outside the limited expressive power of the semantic/pragmatic interpreter or even to expressions that are not strictly correct as referring expressions.

We can improve on this situation by making sure that the user's referring acts are embedded in a discourse—in particular, in an ongoing interaction initiated by the robot and which defines the rules of the game. Thus, if the robot first informs the user what it can perceive in a scene, then the terms and perspectives available for reference are already constrained favorably for the robot's subsequent interpretation. This requires that the robot be in a position to *verbalize* its scene perception. The kind of domain knowledge representation for our newly designed AIBO scenario sketched in figure 6 already goes a considerable way towards this. Standard techniques from the area of natural language generation (e.g., [Horacek, 2001]) work on collections of instances

organized in terms of domain conceptual hierarchies such as the one given here in order to produce natural language descriptions of the requested content. Part of this work involves aggregating the objects present into referring expressions that allow the user to identify what is being described (cf. [Bateman, 1999]). These referring expressions can then already be used to suggest to users particular ways of describing the goal objects of their required instructions.

In a simple case, for example, there may be a scene in which there are several similar cubes within the same spatial sector, but where those cubes differ with respect to some other attribute: two may be red, another blue. Standard aggregation techniques can pick out the differing attributes and use these to determine appropriate referring expressions: thus, we can ascertain that 'the blue cube' is sufficient to identify the one blue cube in question, while 'the red cube' will need further elaboration (e.g., by the projective relations described above). As the situation to be described becomes more complex, correspondingly more complex referring expressions may be produced that are limited in the practical ways already investigated in detail in work such as that of Reiter and Dale (1992) . In particular, aggregration can establish particular groups in the discourse to serve as the relatum in projective expressions of the kind illustrated above. A robot-produced utterance such as "To my left there are two red cubes and one blue cube" introduces in addition to the three objects with the sector 'left' a subgroup consisting of just the two red cubes. Subsequent reference can use this just as the perceptually defined groups were used above: e.g., "the rightmost red cube". Within an interaction, interpretation of this expression can be constrained to the group of red cubes introduced by the robot at that point, rather than referring to the entire set of red cubes possibly available in the scene at large. Reference thus becomes both interactional and situated as is natural in human-human interaction.

In a different scenario, human users may be expected to use spoken language to address the robot, rather than to type their instructions. While such a scenario at first sight is undoubtedly more natural to the users, it raises a range of different problems, such as those occurring when the speech input is not recognised correctly by the system, or those caused by user expectations about the system's facilities. Compared to our scenario, it is by no means clear which kinds of language users would employ if required to talk rather than to type. It is well-known that spoken language differs in many respects from written language; also depending on other situational factors [Biber, 1988]. Further research is needed to explore this and other kinds of variation with regard to the enhanced human-robot interaction scenario with our new robot AIBO which we are now implementing.

7 Conclusion

In this paper, we have described an implemented mobile robot system that follows simple instructions given by its human users. We have investigated empirically the kinds of instructions that users employ and have provided a computational model of these strategies as a level of spatial instruction knowledge representation that interfaces between the linguistic input provided to the robot and the robot's sensing and action component. This implemented version of the system was demonstrated to perform in an

adequate way, but only in a relatively simple set of possible task scenarios. We then briefly sketched a current direction of research in which we are building on the explicit spatial instruction model in order to provide more interactive linguistic behavior. This will feed into a further round of empirical investigation, which will evaluate the effectiveness of the functionalities provided. We have suggested that this is a necessary and beneficial step towards achieving more robust and natural interactional styles between humans and mobile robots.

Acknowledgement

The authors would like to thank Carola Eschenbach, Christian Freksa, Christopher Habel and Tilman Vierhuff for interesting and helpful discussions related to the topic of the paper. We thank Bernd Hildebrandt for constructing the parser. And we would like to thank Jan Oliver Wallgrün, Stefan Dehm, Diedrich Wolter and Jesco von Voss for programming the robot and for supporting the experiments. Also many thanks to Christie Manning for helpful comments on our paper.

References

[Amalberti et al., 1993] Amalberti, R., Carbonell, N., and Falzon, P. (1993). User Representations of Computer Systems in Human–Computer Speech Interaction . *International Journal of Man–Machine Studies*, 38:547–566.

[Bateman, 1999] Bateman, J. A. (1999). Using aggregation for selecting content when generating referring expressions. In *Proceedings of the 37th. Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 127–134, University of Maryland. Association for Computational Linguistics.

[Biber, 1988] Biber, D. (1988). *Variation across speech and writing.* Cambridge University Press, Cambridge.

[Eschenbach, 2001] Eschenbach, C. (2001). Contextual, Functional, and Geometric Features and Projective Terms . In *Proceedings of the 2nd annual language & space workshop: Defining Functional and Spatial Features*, University of Notre Dame.

[Eschenbach et al., 2000] Eschenbach, C., Tschander, T., Habel, C., and Kulik, L. (2000). Lexical Specification of Paths . In Freksa, C., Habel, C., and Wender, K. F., editors, *Spatial Cognition II*, Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin.

[Fischer, 2000] Fischer, K. (2000). What is a situation? In *Proceedings of Gtalog 2000, Fourth Workshop on the Semantics and Pragmatics of Dialogue*, pages 85–92.

[Habel et al., 1999] Habel, C., Hildebrandt, B., and Moratz, R. (1999). Interactive robot navigation based on qualitative spatial representations. In Wachsmuth, I. and Jung, B., editors, *Proceedings Kogwis99*, pages 219–225, St. Augustin. infix.

[Hernández, 1994] Hernández, D. (1994). *Qualitative representation of spatial knowledge*. Lecture Notes in Artificial Intelligence. Springer Verlag, Berlin, Heidelberg, New York.

[Herrmann, 1990] Herrmann, T. (1990). Vor, hinter, rechts und links: das 6h-modell. psychologische studien zum sprachlichen lokalisieren. *Zeitschrift für Literaturwissenschaft und Linguistik*, 78:117–140.

[Herrmann and Grabowski, 1994] Herrmann, T. and Grabowski, J. (1994). *Sprechen: Psychologie der Sprachproduktion.* Spektrum Verlag, Heidelberg.

[Hildebrandt and Eikmeyer, 1999] Hildebrandt, B. and Eikmeyer, H.-J. (1999). *Sprachverarbeitung mit Combinatory Categorial Grammar: Inkrementalität & Effizienz* . SFB 360: Situierte Künstliche Kommunikatoren, Report 99/05, Bielefeld.

[Hildebrandt et al., 1995] Hildebrandt, B., Moratz, R., Rickheit, G., and Sagerer, G. (1995). Integration von bild- und sprachverstehen in einer kognitiven architektur. In *Kognitionswissenschaft*, volume 4, pages 118–128, Berlin. Springer-Verlag.

[Horacek, 2001] Horacek, H. (2001). Textgenerierung. In Carstensen, K.-U., Ebert, C., Endriss, C., Jekat, S., Klabunde, R., and Langer, H., editors, *Computerlinguistik und Sprachtechnologie – Eine Einführung*, pages 331–360. Spektrum Akademischer Verlag, Heidelberg.

[Kummert et al., 1993] Kummert, F., Niemann, H., Prechtel, R., and Sagerer, G. (1993). Control and explanation in a signal understanding environment. *Signal Processing, special issue on 'Intelligent Systems for Signal and Image Understanding'*, 32:111–145.

[Lay et al., 2001] Lay, K., Prassler, E., Dillmann, R., Grunwald, G., Hgele, M., Lawitzky, G., Stopp, A., and von Seelen, W. (2001). MORPHA: Communication and Interaction with Intelligent, Anthropomorphic Robot Assistants. In *International Status Conference: Lead Projects Human-Computer-Interaction*, Saarbruecken, Germany.

[Levelt, 1996] Levelt, W. J. M. (1996). Perspective Taking and Ellipsis in Spatial Descriptions . In Bloom, P., Peterson, M., Nadel, L., and Garrett, M., editors, *Language and Space*, pages 77–109. MIT Press, Cambridge, MA.

[Levinson, 1996] Levinson, S. C. (1996). Frames of Reference and Molyneux"s Question: Crosslinguistic Evidence . In Bloom, P., Peterson, M., Nadel, L., and Garrett, M., editors, *Language and Space*, pages 109–169. MIT Press, Cambridge, MA.

[Moratz, 1997] Moratz, R. (1997). *Visuelle Objekterkennung als kognitive Simulation*. Diski 174. Infix, Sankt Augustin.

[Moratz et al., 1995] Moratz, R., Eikmeyer, H., Hildebrandt, B., Kummert, F., Rickheit, G., and Sagerer, G. (1995). Integrating speech and selective visual perception using a semantic network. *Proc. AAAI-95 Fall Symposium on Computational Models for Integrating Language and Vision*, pages 44–49.

[Moratz et al., 2001] Moratz, R., Fischer, K., and Tenbrink, T. (2001). Cognitive Modeling of Spatial Reference for Human-Robot Interaction. *International Journal on Artificial Intelligence Tools*, 10(4):589–611.

[Moratz and Hildebrandt, 1998] Moratz, R. and Hildebrandt, B. (1998). *Deriving Spatial Goals from Verbal Instructions - A Speech Interface for Robot Navigation* - . SFB 360: Situierte Künstliche Kommunikatoren, Report 98/11, Bielefeld.

[Moratz et al., 2002] Moratz, R., Nebel, B., and Freksa, C. (2002). Qualitative spatial reasoning about relative position: The tradeoff between strong formal properties and successful reasoning about route graphs. *this volume*.

[Neumann and Novak, 1983] Neumann, B. and Novak, H.-J. (1983). Event models for recognition and natural language description of events in real-world image sequences. In *IJCAI 1983*, pages 643–646.

[Niemann et al., 1990] Niemann, H., Sagerer, G., Schröder, S., and Kummert, F. (1990). ERNEST: a semantic network system for pattern understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(9):883–905.

[Oviatt et al., 1998] Oviatt, S., MacEachern, M., and Levow, G.-A. (1998). Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication*, 24:87–110.

[Reiter and Dale, 1992] Reiter, E. and Dale, R. (1992). A fast algorithm for the generation of referring expressions. In *Proceedings of the fifteenth International Conference on Computational Linguistics (COLING-92)*, volume I, pages 232–238, Nantes, France. International Committe on Computational Linguistics.

[Retz-Schmidt, 1988] Retz-Schmidt, G. (1988). Various Views on Spatial Prepositions. *AI Magazine*, 9(2):95–105.

[Schegloff et al., 1977] Schegloff, E., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organisation of repair in conversation. *Language*, 53:361–383.

[Steedman, 1996] Steedman, M. (1996). *Surface Structure and Interpretation*. MIT Press, Cambridge, MA.

[Stopp et al., 1994] Stopp, E., Gapp, K.-P., Herzog, G., Laengle, T., and Lueth, T. C. (1994). Utilizing Spatial Relations for Natural Language Access to an Autonomous Mobile Agent. *Künstliche Intelligenz*, pages 39–50.

[Streit, 2001] Streit, M. (2001). Why Are Multimodal Systems so Difficult to Build? - About the Difference between Deictic Gestures and Direct Manipulation. In Bunt, H. and Beun, R.-J., editors, *Cooperative Multimodal Communication*. Springer-Verlag, Berlin, Heidelberg.

[Wahlster, 2001] Wahlster, W. (2001). SmartKom: Towards Multimodal Dialogues with Anthropomorphic Interface Agents. In *International Status Conference: Lead Projects Human-Computer-Interaction*, Saarbruecken, Germany.

[Wahlster et al., 1983] Wahlster, W., Marburger, H., Jameson, A., and Busemann, S. (1983). Over-answering yes-no questions: Extended responses in a nl interface to a vision system. In *IJCAI 1983*, pages 643–646.