COGNITIVE MODELLING OF SPATIAL REFERENCE FOR HUMAN-ROBOT INTERACTION

REINHARD MORATZ, KERSTIN FISCHER†, and THORA TENBRINK‡
University of Bremen, Center for Computer Studies
Bibliothekstr. 1, 28359 Bremen, Germany
moratz@tzi.de

† *University of Bremen, FB10: Linguistics and Literary Studies*
Postfach 330440, 28334 Bremen, Germany
kerstinf@uni-bremen.de

‡ *University of Hamburg, Department for Informatics*
Vogt-Kölln-Str. 30, 22527 Hamburg, Germany
tenbrink@informatik.uni-hamburg.de

The question addressed in this paper is which types of spatial reference human users employ in the interaction with a robot and how a cognitively adequate model of these strategies can be implemented. In experiments we explored how human users approach an artificial communication partner, which was designed to mimic spatial reference among humans. Our findings show that spatial reference in human-robot interaction differs from natural situations in human-human interaction in several respects. For instance, many users unexpectedly employed fine-grained, path-based, instructions rather than specifying the intended goal object of the action directly. If instructions were not successful, participants created less and less complex descriptions. Those users who did specify the goal object were found to employ those kinds of spatial reference strategies implemented in our computational model. In particular, they exploited the presence of several similar objects by perceiving and referring to them linguistically as a group.

Keywords: robotics, spatial representations, natural language processing

1. Introduction and Motivation

The questions addressed in this paper are which strategies for achieving joint spatial reference human users employ in the communication with a robot, and which results from human-to-human communication can be transferred to human-robot interaction. Such a transfer may be problematic because of the different perceptual capabilities of humans and robots which influence the ways how objects may be referred to, the specific ways of formulating tasks in human-human and human-robot scenarios, restrictions to particular modalities available, and other differences related to the 'naturalness' of the communication. In this paper, we address the peculiarities of human-robot interaction by investigating experimentally how humans interact linguistically with a robot that we designed on the basis of findings on spatial reference in human-to-human communication.

1.1. *Problem statement*

Many tasks in the field of service robotics can profit from a natural language interface. In a typical scenario, a human has to instruct a robot to perform an action on an object. A prominent task for human-robot interaction is then to establish a joint reference to this object. This requires that the scene description created by the robot's object recognition system and the visual system of the human instructor be matched.

Natural language interfaces are oriented to properties of human-to-human communication. Yet, in natural language, the goal object (i.e., the object with which an action is to be performed) is usually specified by the class name of the object, for instance, "please hand me the screwdriver" or "can you pass me the book?". Unfortunately, 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. In contrast, the human sensory apparatus allows for a much greater variety of spatial reference than that of robots since human perceptual abilities are much richer than those of the standard robot. This fact may cause problems for the interaction between human users and the robot, and severe communication problems may occur. 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 human-robot instruction in natural language may be:

> "the small reflecting object on the ground, to the left of the brown box."

because robots have limited perceptual capabilities and need to code the world knowledge explicitly. Thus, while reference to the goal object by means of the category name is often not useful in human-robot interaction, other features of the object, in particular, its spatial arrangement and its position relative to the robot itself can be used for linguistic reference. Alternative strategies are thus, for instance, to refer to objects by means of their color, size, or position.* Qualitative spatial reference may consequently serve as a bridge between metric knowledge required by the robot, and more vague concepts that build the basis for natural linguistic utterances, as suggested by Hernández [1].

Qualitative linguistic spatial reference has been characterized by way of distinguishing various kinds of reference systems (e.g., Levinson [2]). Such classifications are based on naturally occurring scenarios in human-human interaction. For example, 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.

Typical human-robot interaction scenarios, in contrast, may create the need for ways of referring to objects which are less common in natural human-human interaction. In our experimental scenario, for instance, the perspective is reversed: it is not the position of an object that is unknown, but rather the identity of one of several entities, whose positions in space are known, needs to be determined. Thus, the issue at hand is, "Which of these similar objects do you refer to?". This perspective triggers ways of linguistic reference hitherto largely ignored in the literature on spatial reference systems.

Another problematic area is constituted by the restriction to particular modalities in human-robot interaction unknown to the natural communication among humans. For instance, in naturally occuring human-human interaction, in a scenario where both interactants are present, the intended object may be pointed at, without the need for linguistic interaction. The automatic understanding of pointing gestures, however, demands very sophisticated image processing abilities and is not very reliable at the present state of the art, especially where only visual information is available. Thus, the most natural way of achieving reference to a goal object in joint attention scenarios may not be available in human-robot interaction.

In addition, human speakers in human-robot scenarios may be to a high degree uninformed about the robot's features, and thus unsure about how they should address the system [3]. This insecurity may be a reason for a number of linguistic strategies peculiar to human-robot communication.

To conclude, human-robot interaction differs from natural human-to-human communication since the perceptual capabilities of robots, particularly their incapability to distinguish objects on the basis of non-spatial features, such as functional categories, colours, sizes, etc., are much poorer than those of human interlocutors. For this reason it may more often be necessary for robots than it is for humans to rely on the objects' relative position in space in order to establish reference. Moreover, as outlined above, typical human-robot interaction sce-

*Being interested in spatial reference rather than in aspects of categorization etc., in this paper we focus particularly on the use of positional information for reference to objects in human-robot interaction.

narios differ in kind from most situations in natural human-human interaction. Thus, it is by no means self-evident how speakers react to a robot instruction task in an open scenario such as the one we employ in the experiments to be reported on in this paper.

1.2. *Approach*

The methodology used in this study is to have human users interact with a robot which was designed on the basis of cognitive adequacy regarding what is known about spatial reference in human-to-human communication. The starting point is the hypothesis that in general, users and the system interactively achieve a common mode of communication. Speakers are expected to try out as many different strategies as necessary for a successful interaction with their communication partner, similar to real world interactions with people whose capabilities are difficult to estimate, such as children, handicaped, or foreigners. Furthermore, speakers accomodate to the listeners due to "speakers' *intense* desires for social approval, interpersonal affiliation or group identification" on the one hand and the wish "to increase evaluations of competence, culture or control, and to emphasize social distance" on the other [4]. Regarding human-computer interaction, Amalberti et al. [5] have shown that while speakers initially approach human and artificial communication partners very differently, the two types of linguistic behaviour become increasingly similar over time if the linguistic behaviour by their communication partner is actually identical. That is, in these experiments, the output was manipulated by a human 'wizard' who did not know whether the participants were instructed to be talking to a machine or to a human communication partner. Thus, both those speakers who believed to be talking to a machine and those who were informed that they were talking to a human being were confronted with a similar linguistic behaviour by the wizard. The fact that the differences between the two speaker groups decreased points to adaptation processes to the linguistic behaviour of the communication partner.

Similarly, in error resolution contexts, speakers' adaptations which are intended to increase understandability can be found as well [6,7,8]. Moreover, speakers have been found to be extremely patient in what they endure with malfunctioning artificial communicators [9]. Thus, in human-robot interaction, speakers are expected to adapt to their artificial communication partner on the basis of its linguistic and behavioural output and thereby allowing us to see the traces of their ideas about what the communicative failure could be caused by. In particular, it can be expected that if attempts to instruct the robot turn out to be unsuccessful, users change their strategy and try another one, for instance, a different type of spatial reference, a different perspective or different lexical material. The experiments then provide us with a rich overview of the strategies speakers preferably use in the interaction with a robot.

Correspondingly, our procedure was firstly to design a computational model of spatial reference for our robot on the basis of psycholinguistic results on spatial reference among humans, supplemented by scenario-specific assumptions. Secondly, in a set of experiments carried out for the present paper the strategies were identified that speakers employed in the interaction with the robot to achieve joint spatial reference. The users' behaviour was then analysed linguistically.

2. Spatial Reference Systems

To set up a cognitively adequate model of verbal strategies of spatial reference, results from psychology and psycholinguistics on spatial expressions in human-to-human communication were integrated. We used the surveys presented by Levinson [2], Herrmann [10], and Levelt [11], and we extended them to be able to refer to one object in a group of similar objects, as we will show in the second subsection. In the last subsection of this section, we introduce our computational model.

2.1. *Intrinsic, relative, and absolute reference systems*

Previous research on reference systems for spatial descriptions has led to the identification of three different reference systems with three variations each, dependent on whether the speaker, the hearer, or a third entity serves as the origin of the perspective employed. The three different options are labeled by Levinson [2] as *intrinsic*, *relative*, and *absolute.*

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 [10] refers to this option as *two-point localisation.*

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 inbuilt features of the relatum. Thus, the stone (= referent) may be situated to the left of the house (= 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.

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* [2]) serve as anchor directions. Thus, the stone may be to the north of the speaker, the hearer, or the house. 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.

2.2. *Reference to groups of objects*

The above classification of reference systems has proved useful for identifying one object's position in relation to one or two other entities. However, such a distinction does not shed any light on how speakers act in a situation where two or more objects of the same natural kind are involved. In such a scenario, objects may be classified (i.e. perceptually grouped) into and referred to as groups rather than individual objects. How one of the objects is picked out linguistically and identified unambiguously is a new question.

We view the group as a special kind of *relatum* when the position of one of its objects (the referent) is referred to by determining its position relative to the rest of the group. In contrast to the classification outlined above, the relatum is not just one entity (e.g. the house), but there are several similar ones which may or may not possess intrinsic features feasible for spatial reference. Moreover, the referent is a part of the relatum, and may be linguistically indistinguishable from the rest of the relatum, apart from the spatial position.

Different subtypes of relata lead to different kinds of consequences for each of the three types of reference systems (intrinsic, relative, and absolute), as each type makes different use of the object identified as relatum.

In intrinsic reference systems, the presence of a relatum with intrinsic features (other than referent, speaker, or hearer) renders the speaker's or hearer's position irrelevant. If groups of objects serve as a 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 into one direction one could refer to "the one who walks in the front of the group".

In relative reference systems, a stone (the referent) may be referred to in relation to a single entity, e.g., a house, as relatum. Similarly, the stone may be referred to in relation to a group of other stones. Then, 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, the stone may be to the north of the rest of the group, independent of the speaker's or hearer's point of view. Thus, "the stone which is furthest to the north" would be a suitable and unambiguous expression.

2.3. *A computational model for a spatial human-robot interaction scenario*

The computational model of spatial reference implemented rests on the following assumptions. First of all, although humans generally use their own point of view in spatial reference, they tend to adopt their listener's perspective under certain circumstances, such as when talking to children, handicaped, or when action by the listener is involved [12]. As our scenario both involved action by the listener, and an interlocutor with different cognitive abilities than the human speaker, and because our robot was, in fact, unable to perceive the speaker at all, we excluded intrinsic and relative reference systems employing either the speaker or the salient object as origin. Secondly, the absence of intrinsic features in the group of similar objects excludes the usage of intrinsic reference systems employing the group as both relatum and origin. Thirdly, the speakers would refrain from using absolute reference systems as these are rarely used in natural human-human interaction in Western culture in indoor scenarios.

Accordingly, three different kinds of linguistic spatial reference offered themselves for communication in our scenario:

First, the speakers could employ an intrinsic reference system using the robot's position as both relatum and origin. In this case, they would specify the object's position relative to the robot's front. Secondly, they could refer to a salient object, if available, as relatum in a relative reference system. Then, they would specify the object's position relative to the salient object from the robot's point of view. Finally, they could refer to the group as relatum in a relative reference system. In this case, they would specify the object's position relative to the rest of the group from the robot's point of view.

We equipped the robot with a computational model representing the different kinds of reference systems to parse linguistic references according to these three options and to handle the corresponding instructions. The model can be outlined as follows.

To model reference systems that take the robot's point of view as origin, as is the case in all three of the expected kinds of reference systems, 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 1), which may be the robot itself, the group of objects, or other salient objects.

The partitioning into a left and a right half-plane is a sensible model for the directions "left of" and "right of" relative to the relatum. The dichotomy front/back is modelled similarly by using another axis orthogonal to the ref-

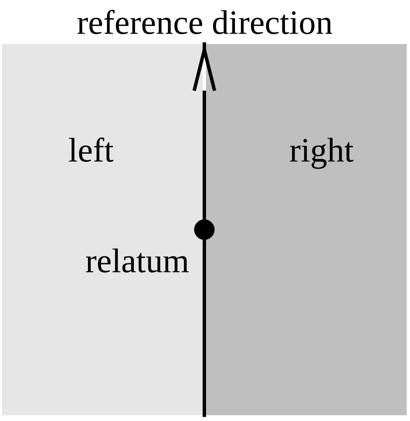


Fig. 1. Relatum and reference direction.

erence axis (see figure 2). However, this representation does only apply 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 [12]. The result is a qualitative distinction, as suggested, for instance, by Freksa [13].

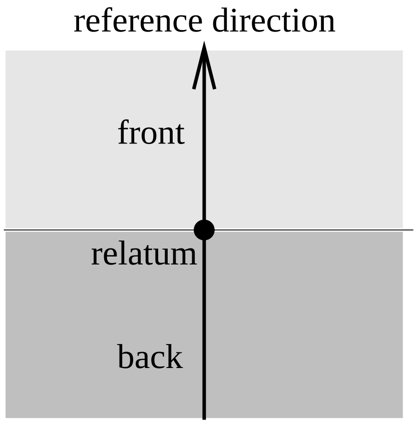


Fig. 2. Front-back dichotomy.

If the robot is chosen as relatum, the reference direction is naturally given by its view direction. The view direction of the robot is its symmetry axis and therefore a salient structure to be observed by the instructor.

If the referent is closer to another salient object than to the robot, this object is assumed to be a convenient relatum. In this case, the reference direction is given by the directed straight line from the robot to the relatum (see 3). In this variant of relative localisation, the “in front of” sector is directed towards the robot.

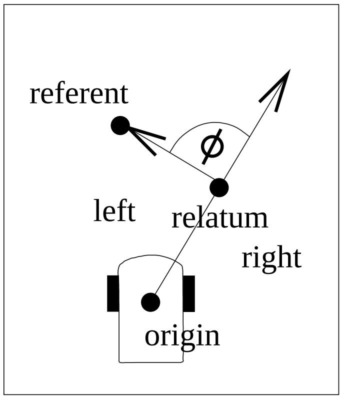


Fig. 3. Relative reference model.

To define the partitions formally we refer to the angle ϕ between the reference direction and the directed straight line from the relatum to the referent (see figure 3).

$$\begin{aligned}
referent \text{ front } relatum &:= -\pi/2 < \phi < \pi/2 \\
referent \text{ left } relatum &:= 0 < \phi < \pi \\
referent \text{ back } relatum &:= \pi/2 < \phi < 3/2\pi \\
referent \text{ right } relatum &:= -\pi < \phi < 0 \\
referent \text{ left front } relatum &:= 0 < \phi < \pi/2 \\
referent \text{ left back } relatum &:= \pi/2 < \phi < \pi \\
referent \text{ right front } relatum &:= -\pi/2 < \phi < 0 \\
referent \text{ right back } relatum &:= -\pi < \phi < -\pi/2
\end{aligned}$$

For a group of similar objects, the centroid of the group can be treated as the relatum. Analogous to the salient object reference model, the reference direction is given by the directed straight line from the robot center to the group centroid. Thus, in this variant of relative localisation involving groups of similar

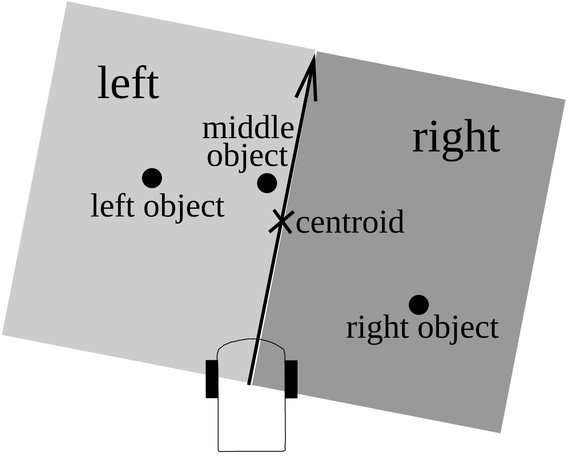


Fig. 4. Group-based Reference.

objects, the group centroid serves as virtual relatum. The object closest to the group centroid can be referred to as the "middle object" (see figure 4).

3. The Natural Language Controlled Robot System

Our goals in realizing this system are twofold, scientifically and application oriented: On the one hand, the system serves as the experimental means to test theoretical hypotheses about the interaction of underdetermined conceptual representations and sensorical input of spatial environments, on the other hand, natural language interfaces, which allow natural instructions, will make robot applications accessible for non-expert users.

The architecture of the system is described in detail in Habel et. al. [14]. We summarize here the main properties of the system's components. The following components interact in the system: the syntactic component, the semantic component, the spatial reasoning component, and the sensing and action component (see figure 5).

The *syntactic component* is based on Combinatory Categorial Grammar (CCG), which has been developed by Steedman [15]. In addition to the domain-dependent development of the lexical categorial system, an adaptation of the grammar rules to German and to the "spurious ambiguity" of CCG was necessary. Specific categorial rules were derived from the original CCG rules to make incremental and efficient processing possible [16].[†]

From feature-value structures, which are the output of the syntactic component, the *semantic component* produces underdetermined propositional rep-

[†] The syntactic component was developed as part of SFB 360 at the University of Bielefeld.

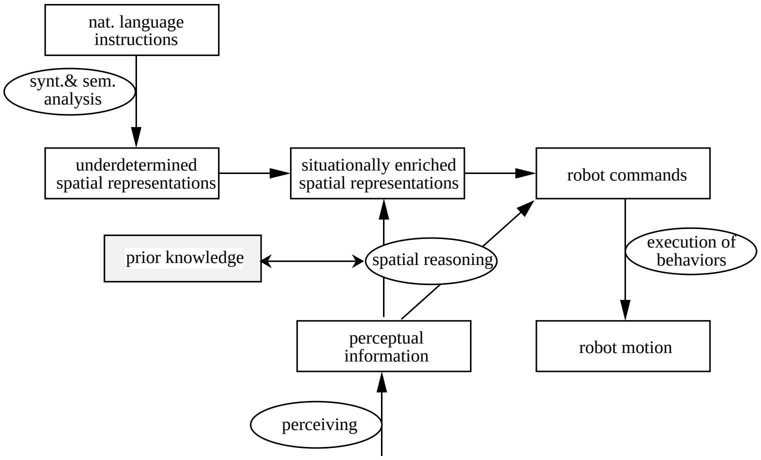


Fig. 5. Coarse architecture of a NL-instructable robot: modules and representations (from Habel et. al. [14]).

resentations of the spatial domain. The component also implements the computational model of projective relations described in section 2.3. It 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. Therefore the robot's perception-based spatial representation of the environment is partitioned using a cell decomposition approach (see Habel et. al. [14] for details). The centroids of the cells are connected with straight lines. The graph which has the centroids as nodes can be viewed as a qualitative map of the environment. After the robot has built up a qualitative map, the path finding task from start to goal location is reduced to a simple graph search algorithm. This method is similar to the one described by Thrun [17].

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 a sensory equipment resembling human sensorial capabilities [18]. 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 6). The images are processed with region-based object recognition [18]. The spatial arrangement of these regions is delivered to the spatial reference component as a qualitative relational description. The *behavior*

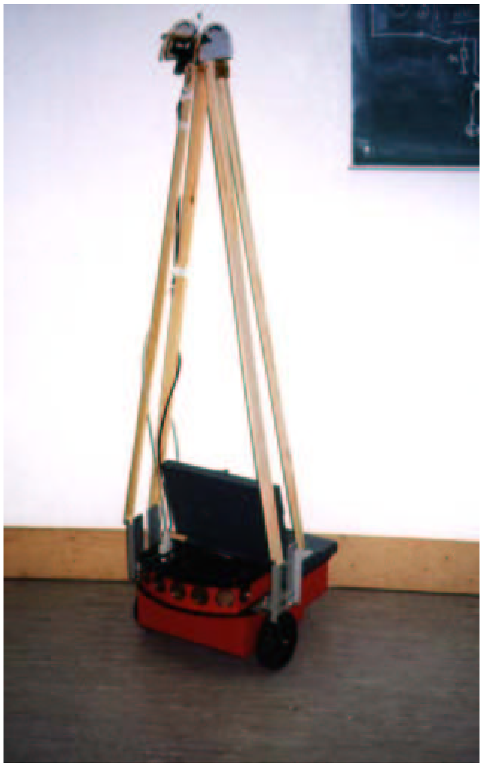


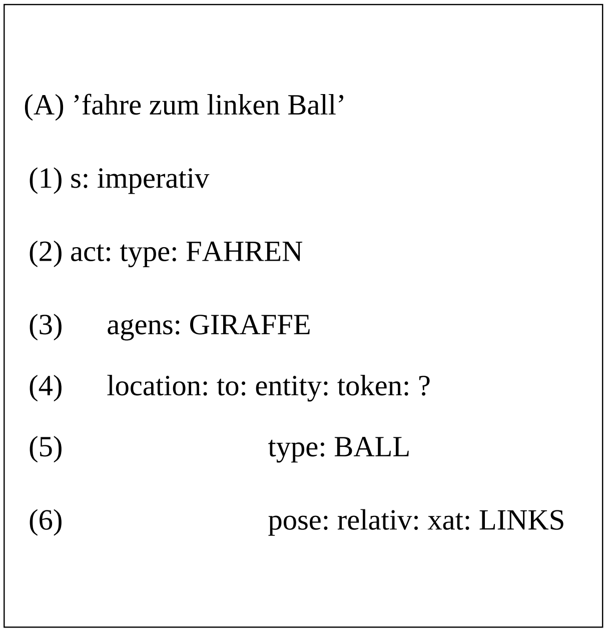
Fig. 6. Our robot GIRAFFE.

execution subcomponent manages the control of the mobile robot (Pioneer 1). This subcomponent leads the robot to perform turnings and straight movements as basic motoric actions. These actions are carried out as 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 left ball")[‡] has the semantic interpretation which is shown in figure 7.

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



(A) 'fahre zum linken Ball'

(1) s: imperativ

(2) act: type: FAHREN

(3) agens: GIRAFFE

(4) location: to: entity: token: ?

(5) type: BALL

(6) pose: relativ: xat: LINKS

Fig. 7. Semantic interpretation.

Now an object that denotes "the left 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. Then, the spatial reasoning component needs to find some suitable space for movement in order to establish a qualitative route graph. This procedure can be illustrated by an example scenario (see figure 8) involving four cubes.

As the background of the image is not varicolored, it has a low saturation value. Therefore a simple color segmentation can dissociate the coloured cubes from the background. The contour of the objects is used to specify a geometrical feature vector which is used to classify the objects. The localisation of the objects assumes that the objects are perceived from an elevated position and are situated considerably below the height of the camera. The centroid of the area is identified. That way an affine transformation can approximate a plan view. This corresponds to the projection $p_{\mathcal{D}}(O)$ of an object O onto the plane (see subsection 2.3. Now we have a good approximation of the point-like repre-

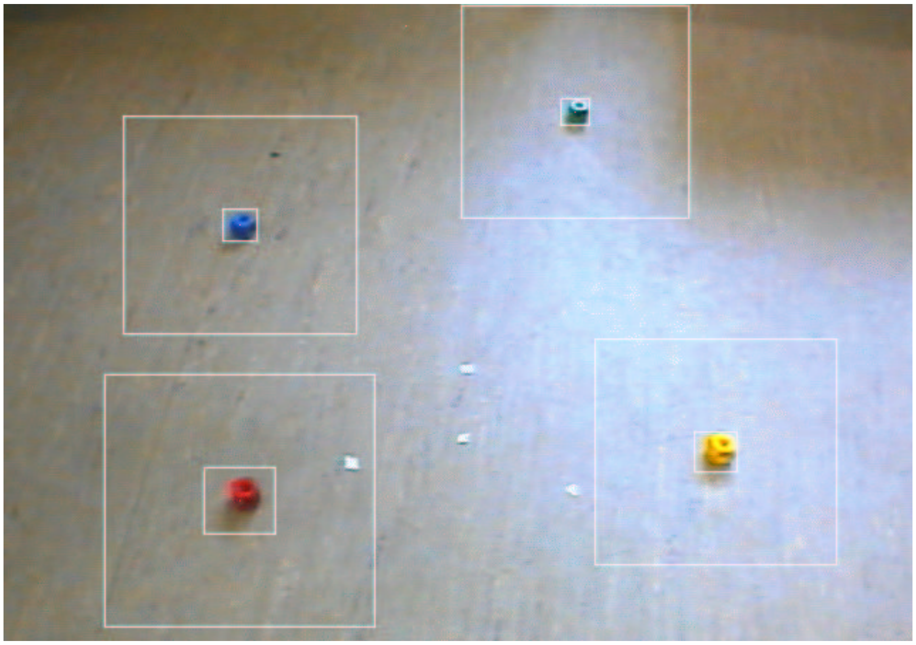


Fig. 8. Object recognition and localisation.

sentation O' of the object O: $O' = \mu(p_{\mathcal{D}}(O))$. The spatial arrangement of the pointlike objects is delivered to the spatial reasoning component as a qualitative relational description.

This spatial reasoning component conceptualizes the robot as an extended circular object. This is already a good approximation for our Pioneer 1 robot. The algorithm only takes into account as suitable the space which has the radius of the robot as minimal distance to the objects (simplified to a square for point-like objects, see figure 9). This space is partitioned into convex cells. The boundaries of the cells are marked off in red in the image. The centroids of the cells are interconnected between neighboring cells, thus generating the qualitative route graph.

Now, a path to the cell containing the goal can be found by means of a simple route graph search whenever the semantic component selects a goal. Within this cell, a direct goal approach will be invoked.

4. Experiments

We utilized our robot equipped with the described functionalities in order to investigate the following questions:

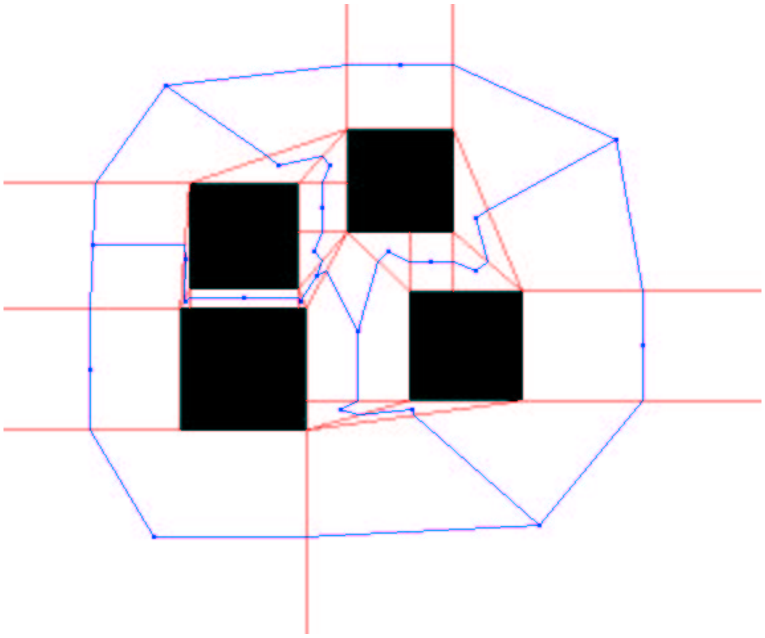


Fig. 9. Cell decomposition and route graph.

- Given the robot whose design was based on the criterion of cognitive adequacy for spatial reference in human-to-human communication, how do human users achieve joint reference in spatial human-robot interaction?
- Which spatial reference systems do users employ and how are these expressed linguistically?
- How do users distinguish between individuals of groups of similar objects?

Furthermore, the experiment served to evaluate our system's functionalities. In order to carry out the experiments in human-robot interaction, we set up a scenario in which technologically naive users were asked to instruct our robot *Giraffe* (see figure 6) to move to one of several roughly similar objects.

4.1. *Experimental design*

In order to answer the above questions, a test scenario was developed in which the user's task was to make the robot move to particular locations pointed at by the experimenter; pointing was used in order to avoid the prompting of verbal expression or pictures of the scene which would impose a particular perspective, for example, the view from above. Users were instructed to use natural language

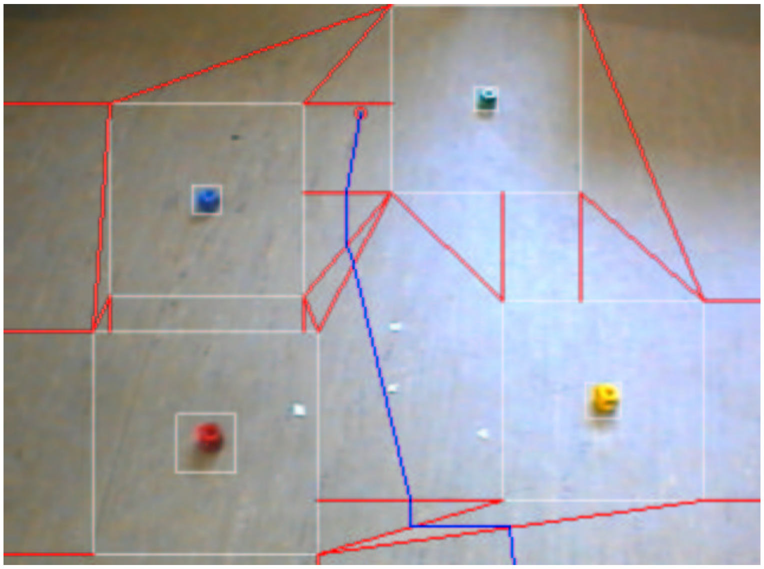


Fig. 10. Path planning.

sentences typed into a computer to move the robot; they were seated in front of a computer in which they typed their instructions. When they turned around, they perceived a scene in which, for instance, a number of cubes were placed on the floor together with the robot, which was placed in a 90 degree angle or opposite of the participant, as shown in figure 11. The fixed setting allows us the analysis of the point of view the participant was taking, depending on the instruction employed. The arrangements of the cubes were 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.

As outlined above, the robot could 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”. Thus, human users who were not successful from the start were challenged to try out many different kinds of spatial instructions to enable the robot to identify the intended aim.

15 different participants carried out an average of 30 attempts to move the robot within about 30 minutes time each. Their sentences were protocolled, and their verbal behaviour during the experiments was recorded in order to capture self-talk in which speakers announced their strategies or their ideas about what was going wrong. After the experiments, participants were asked as to what they believed the robot could and could not understand, which strategies they

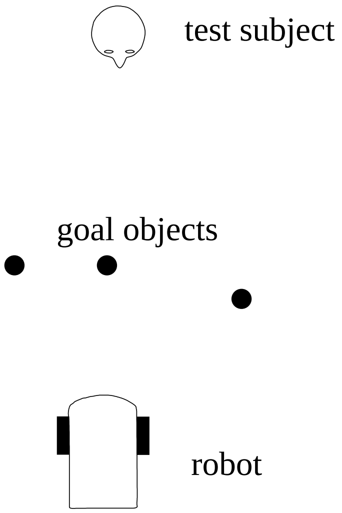


Fig. 11. The setting of the experiment.

believed to be non-successful, and whether their beliefs about the robot had changed during the interaction. Altogether 476 instructions were elicited.

4.2. ***Experimental results***

We focus on three areas of special interest addressed in the current experiment that may be compared to findings from human-to-human communication: the strategy of instruction, the perspective employed, and the types of reference systems used.

4.2.1. *Instructional strategy*

For object names, participants used mainly basic level categories, such as *cube* or *block.* Only one user employed the more abstract category *object.* This is fully consistent with our hypotheses based on findings from natural human-to-human communication. However, a surprising finding is that only half of the participants used the goal-object describing strategy expected (see section 2.1.). Examples of this strategy are *fahr bis zum rechten Würfel* [**drive up to the right cube**], *fahr zu dem Klotz, der vor Dir liegt* [**drive to the block which is in front of you**], *geh zu dem vorderen Würfel* [**walk to the front cube**]. Since this strategy was the one expected and implemented, these instructions were usually successful, unless there were orthographic, lexical, or syntactic problems. In such cases, the participants usually tried out path-naming strategies next; if successful, they sticked to the goal-naming strategy in later instructions.

The other half of the participants, however, instead of naming the goal object itself, first tried out a much simpler strategy, decomposing the action by describing the path the robot had to take, for instance, *fahr 1 Meter geradeaus* [**drive 1 meter ahead**], *rolle ein wenig nach vorn* [**roll a bit forward**], *fahre nach Norden von Dir aus gesehen* [**drive north from your point of view**], *links* [**left**], *los Du lahme Kiste vorwärts nach rechts* [**come on you lame box ahead to the right**], *bewege Dich Richtung Schrank* [**move in the direction of the wardrobe**].

If the path descriptions did not work, the participants did not try out a description of the goal object, which the robot would have understood. Instead, they used descriptions of movements, for instance, *fahre* [**drive**], *bewege Dich mit einer positiven Geschwindigkeit in irgendeine Richtung* [**move with positive speed in some direction**], *sitz* [**sit**], *spring* [**jump**], *Drehung!* [**turn!**]. Some participants who had used this strategy, employed afterwards a fourth one, namely to specify the instrumental actions necessary for such movement, for example: *drehe Deine hinteren Rollen* [**turn your rear wheels**] or *Motor an* [**engine on**].

The path-description strategy seemed very natural to the participants, and they were on the whole quite desperate to find that this strategy did not work with the robot. Because the situation was so depressing for these participants, they were sometimes hinted at the goal-naming strategy by the experimenter, but even if prompted to try goal descriptions, they did only reluctantly change their strategy in this respect.[§]

Thus, a fixed order of instructional strategies becomes apparent: If goal descriptions were unsuccessful (for other, possibly lexical or orthographic, reasons), users tried out path descriptions, or they started off with path descriptions immediately. If path descriptions turned out unsuccessful, participants employed descriptions of movements. If these proved insufficient, users attempted to instruct the robot by describing actions instrumental to movement in general.

Regarding the syntactic structures used, it is noteworthy that speakers made extensive use of imperatives, a linguistic construction which is rarely used in human-to-human conversation due to politeness reasons [19]. Besides using imperatives, speakers often used infinitives, as it has been reported about people talking to foreigners [20].

Regarding other linguistic choices, users displayed the willingness to adapt to the robot as expected, for instance, by varying lexical items (e.g. drive, move, walk, go), by reducing morphological complexity and by attending to orthographic or stylistic variabilities (e.g. formal vs. informal versions of forms of address or the orthographic representation of imperatives).

[§] The prompting by the experimenters was, of course, recorded and considered in the linguistic analysis.

4.2.2. *Point of view*

As described above, we expected a tendency for human users to adopt their interlocutor's point of view because of the robot's different cognitive abilities, and because of the involvement of action by the listener. This expectation was exceeded by far, as users *consistently* used the robot's perspective. That they attended to the origin as a relevant information furthermore becomes clear from verbal statements recorded during the interaction: In one experiment, the user's first question was where the front of the robot was, that is, the participant found this information to be relevant for the production of an instruction. In another experiment the user had firstly taken the robot's point of view as origin, but due to some other mistake, the instruction was not carried out successfully. The user then announced that she had found out that the robot was using her perspective after all, and she tried out the next instruction accordingly. Thus, human users attend to the point of view as an important informational resource, while at the same time they consistently take the robot's perspective, as long as they do not have evidence that this could not be the right strategy. This finding confirms our hypothesis that users are extremely motivated to achieve successful communication even though this may mean to afford higher mental costs [21,12]. However, for the most part, in goal-based instructions users did not state explicitly that they assumed the robot's point of view, although they consistently used it. This failure to be explicit and unambiguous is consistent with previous findings [12]; especially in cases where the adopted point of view is not unequivocal it can lead to misunderstandings.

4.2.3. *Types of reference systems*

As noted above, there are different strategies participants employed to achieve spatial reference. In this section, we analyse the types of reference systems employed in the goal-naming strategy. Out of the 183 linguistic instructions 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. This instruction has the form of an imperative combined with a locative directional adjunct which specifies the relative position of the cube in the group, for instance, *Fahr zum linken Würfel* [**Drive to the left cube**]. In the common structure, the lexical slots of the verb and the object label were varied, as well as the positional adjective, yielding "mittleren" [**mid**], "hinteren" [**back**], "vorderen" [**front**] as well as "linken" [**left**]. Interestingly, there was one case involving a linguistic form which does not exist in standard German, namely, *linkesten* [**leftmost**]. This usage demonstrates the user's desire to make the fact explicit that she was referring to one specific object within a group of cubes which all might be referred to as being left, i.e., on the left hand-side. In addition to the 69 group-based instructions described so far, nine more utterances used the group for reference by referring to the relative distance of the cubes, as in *Fahre zum weiter entfernten Klotz* [**Drive to the farthest cube**].

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 that specifies 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 utterances (out 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 here, the speakers could not use an expression like "to your right", or "move forward", without assuming a front.

Interestingly, when using path descriptions instead of referring to the goal, several participants employed absolute reference systems, such as *Gehe nach Norden* [**Go to the North**]. As we predicted, however, this type of reference system did not occur in the goal-based instructions.

To sum up, among the goal object describing instructions, group-based reference occurred most frequently. If a salient object was present, it was often used as a relatum. Furthermore, participants made use of the robot's features, implying the robot's front in using intrinsic reference systems. Thus, our expectations with regard to the three kinds of reference systems available for communication in our scenario (see section 2.3.) were fulfilled, yielding successful communication with the robot since the computational model reflected these three kinds of reference systems. Linguistically, especially in group-based reference the instructions were presented often with a similar syntactic structure, i.e., an imperative with an adjunct. In other spatial reference systems, the expressions on the linguistic surface were more varied and generally involved greater linguistic complexity.

5. Conclusion

The empirical research reported on in this paper was meant to show which strategies of spatial reference human users employ in the interaction with a robot and how these may differ from spatial reference found for spatial reference among humans. To reach this aim, firstly a robot was implemented as a cognitively adequate model of spatial reference in human-to-human communication. Experiments in human-robot interaction were then carried out to detect the strategies human users employ to approach an artificial communication partner, yielding both expected and unexpected results. Unexpectedly, about half of the participants did not name or describe the goal object itself, but they described paths how to get there, decomposing the action for the robot. As we did not expect such strategies, we did not represent them in our computational model, which led to communicative failure.

While the participants demonstrated their eagerness to establish successful communication by various strategies such as adapting to the robot's assumed linguistic and perceptual abilities, they were remarkably reluctant to depart from the hierarchy demonstrated in section 4.2.1., NAmely, to begin with either goal or path instructions and then in the case of failure proceed to more detailed instructions. Possibly, the underlying assumption is that goal-based instructions are too complicated for a robot, and that path or movement descriptions are more basic and therefore less difficult. Such assumptions need to be accounted for in order to achieve successful human-robot communication. There was also a high consistency between users in the usage of imperatives, which is not often found in natural human-to-human communication as it points to a highly unequal relationship between the communication partners. These findings show that designers of human-robot interaction systems cannot solely rely on results from human-to-human communication.

As another result, participants were consistently found to use the communication partner's perspective in deciding, for instance, what is left or right or what is front and back. While in natural human-to-human communication a *tendency* has been noted for humans to take their interaction partner's point of view in certain situations, in our human-robot interaction scenario there seemed to be hardly any doubt about the perspective to be taken. In order to achieve joint reference in spatial human-robot interaction, it is thus necessary to take into account findings regarding which kinds of reference systems seem natural and unambiguous to human users, so that the robot can be enabled to neglect unlikely interpretations even if they are logically possible, such as interpretations that employ the speaker's point of view. Consequently it does not suffice to rely on psycholinguistic findings based on human-to-human interaction, because humans display very specific assumptions regarding human-robot interaction which need to be accounted for, and because human-robot interaction scenarios may differ in kind from natural human-to-human interaction situations.

Another result of our research is the significance of groups of similar objects that has been largely neglected in the literature on spatial reference so far. It seems, however, that where similar objects are involved, reference to the group as a whole is an easy and natural way of specifying one's goal. In our setting, participants assumed reference to be effective by distinguishing the goal object from the rest of the group by its spatial position. Thus, for our setting involving roughly similar objects, the concept of a *group of objects* turned out to be central. Previously no computational model existed for group-based reference. We therefore developed a simple model to achieve a representation of the concept of groups of objects. Our experimental results correspond to our assumptions regarding human users' linguistic choices for group-based spatial reference, and they point to ways in which a cognitively adequate model of spatial reference peculiar of human-robot interaction can be extended.

Regarding the design of the current robot, an integration of different modalities, that is, linguistic input, perception and action, was achieved. With respect to goal-centered instructions, our computational model of projective relations was demonstrated to perform in a cognitively adequate way: on the one hand, the users overwhelmingly employed the robot's perspective. On the other hand, most of the spatial reference systems employed corresponded directly to those implemented such that successful communication was achieved.

Acknowledgement

The authors would like to thank John Bateman and Paul Lee for very helpful comments on an earlier version of this paper. Also, we thank Christian Freksa, Christopher Habel and Carola Eschenbach 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.

References

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

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

[3] K. Fischer. How much common ground do we need for speaking? In Peter Kühnlein, Hannes Rieser, and Henk Zeevat, editors, *Proceedings of the 5th Workshop on Formal Semantics and Pragmatics of Dialogue, Bi-Dialog 2001, Bielefeld, June 14-16th, 2001*, pages 313–320, 2001.

[4] H. Giles and A. Williams. Accommodating hypercorrection: A communication model. *Language and Communication*, 12(3/4):343–356, 1992.

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

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

[7] S. Oviatt, J. Bernard, G.-A. Levow. Linguistic Adaptations during Spoken and Multimodal Error Resolution. *Language and Speech*, 41(3-4):419–442, 1998.

[8] G.-A. Levow. Characterizing and Recognizing Spoken Corrections in Human-Computer Dialogue. *Proceedings of Coling/ACL '98*, 1998.

[9] K. Fischer. Repeats, reformulations, and emotional speech: Evidence for the design of human-computer speech interfaces. In Hans-Jörg Bullinger and Jürgen Ziegler, editors, *Human-Computer Interaction: Ergonomics and User Interfaces, Volume 1 of the Proceedings of the 8th International Conferen ce on Human-Computer Interaction, Munich, Germany*, pages 560–565. Lawrence Erlbaum Ass., London, 1999.

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

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

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

[13] K. Zimmermann and C. Freksa. Qualitative spatial reasoning using orientation, distance, and path knowledge. *Applied Intelligence*, 6:49–58, 1996.

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

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

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

[17] S. Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artificial Intelligence*, 99:21–71, 1998.

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

[19] P. Brown and S. Levinson. *Politeness. Some Universals in Language Usage.* Cambridge University Press, 2nd (original 1978) edition, 1987.

[20] J. Roche. *Xenolekte. Struktur und Variation im Deutsch gegenüber Ausländern.* Berlin, New York: de Gruyter, 1989.

[21] M. H. Long. Adaption an den Lerner. Die Aushandlung verstehbarer Eingabe in Gesprächen zwischen muttersprachlichen Sprechern und Lernern. *Zeitschrift für Literaturwissenschaft und Linguistik*, 12:100–119, 1982.