# Cognitively Adequate Modelling of Spatial Reference in Human-Robot Interaction *

Reinhard Moratz and Kerstin Fischer
University of Hamburg, Department for Informatics
Vogt-Kölln-Str. 30, 22527 Hamburg
moratz, kfischer@informatik.uni-hamburg.de

## Abstract

*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 adequat model of these strategies can be implemented. Experiments in human-robot interaction were carried out which show how human users approach an artificial communication partner which was designed on the basis of empirical findings regarding spatial references among humans. Results are considerable differences in the strategies speakers employ to achieve spatial reference in human-robot interaction and in natural communication.*

## 1 Introduction and Motivation

Many tasks in the field of service robotics can profit from a natural language interface. In a typical scenario, there is an object and an action to be performed with it. In natural language, and this has partly also been found to be true in the data of human-robot interaction elicited for this paper, the goal is usually specified by the class name of the object. In an open scenario, however, in which the robot has no detailed *a priori* knowledge about all of the relevant objects, the current state of the art does not allow correct object categorization. Alternatively, objects can also be referred to by their color, size, or position. This paper addresses the use of positional information for reference to objects in human-robot interaction.

Natural language interfaces are oriented to properties of human-to-human communication. However, the human sensory apparatus allows for different kinds of spatial reference since the perceptual abilities of the standard robot are much poorer than human capabilities. This fact may cause problems for the interaction between human users and the robot; thus, while reference used in interpersonal communication would often make use of the object category name:

> " the key on the floor,"

a human-robot instruction in natural language should be similar to the following, given the perceptual constraints and the necessity of explicit coding of world knowledge for robots:

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

The question asked in this paper is now: Which strategies do human users actually employ in the communication with a robot? The methodology used here 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. Therefore, speakers are expected to try out as many different strategies as necessary for a successful interaction with their communication partner, as it happens in real world communication as well, for instance, in the interaction with people whose capabilities are difficult to estimate, such as children, handicaped, or foreigners. Furthermore, speakers have been found to accomodate to each other 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 [3]. Regarding human-computer interaction, Amalberti et al. [1] have shown that while speakers initially approach human and artificial communication partners very differently, if the linguistic behaviour by that partner is identical, the two types of linguistic behaviour become more and more similar after

*This work was supported partly by the *Spatial Inference* project of the DFG priority program on Spatial Cognition, under grant Fr 806/7-2 for Christian Freksa and partly by the BMBF under grant number 01 IV 701 F7.

some time.[1] In error resolution contexts, speakers' adaptations which are intended to increase understandability can be found as well [14, 13, 10]. And, finally, speakers have been found to be extremely patient in what they endure with malfunctioning artificial communicators [2]. Thus, it can be expected that in human-robot interaction, speakers adapt to their artificial communication partner on the basis of its linguistic and behavioural output.

Correspondingly, our procedure was firstly to design a robot on the basis of what is known about spatial reference among humans, secondly, it was identified in a set of experiments carried out for the present paper what the strategies are which speakers employ in the interaction with the robot; since results from other areas of communication point to the fact that human users adapt very much to their communication partners, it was expected that if attempts to instruct the robot turned out to be unsuccessful, users would change their strategy and try another one, for instance, a different type of spatial reference, a different perspective or different lexical material, so that the experiments would provide us with a rich overview of the strategies speakers prefarably use in the interaction with a robot. To see which linguistic strategies users employ in the interaction and how they develop, thus experiments were carried out which show how humans attempt to achieve spatial reference; their communication partner, the robot, was initially designed on the basis of findings from previous research. The results of the experiments point to the ways in which a cognitively adequate model of spatial reference peculiar of human-robot interaction can be extended.

## 2 Spatial References

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 refer primarily to the surveys presented by Levinson [9] and by Levelt [8]. Spatial reference can be communicated by humans either through language or through pointing gestures.

Verbal expressions typically contain projective relations (e.g. "left") that are dependent on a specific perspective or view point [6]. Projective relations use a reference object, a reference direction and qualitative angular sectors as the directional component to specify regions in which the refered object lies. Reference objects can be the speaker, the listener, or other, explicitly referred to, salient objects (e.g. "From my point of view, the coin is to the right of the ball"). In the communication between humans, the speaker typically uses his own direction of view as reference direction; only in some situations, for instance, in the communication with children [11], the speaker uses the listener's reference system in order to simplify reference resolution for the listener [6].

The principles of spatial reference in human-to-human communication described are now applied to constructing a computational model of spatial reference, taking into account the perceptive capability of the robot.

### 2.1 A Computational Model for Spatial References

The model for processing spatial references developed relies only marginally on the (limited) perceptual component of the robot. We do not assume that the robot can perceive the instructor visually. In this case, we only examine those scenarios that use the robot as one fixed point in determining the direction of a reference system. The psychological results mentioned above give reason to assume that users will support the robot by using its reference direction when they give instructions. This assumption was verified in the experimental phase.

#### 2.1.1 The Computational Model of Reference Systems for Projective Relations

To model robot-centered reference systems, all objects are arranged in a bird's-eye view. 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 $\mu$ of this area can be used as point-like representation $O'$ of the object $O$: $O' = \mu(p_{\mathcal{D}}(O))$.

For a reference system, a reference object $RO'$ and a reference axis $\vec{r}$ are required. This reference axis is a directed line through the point $RO'$. These geometric elements partition the plane into a left and a right half-plane (see figure 1).

We use two different partitions of the plane, one for interpretation and one for generation. This is motivated by the desire that the acceptor model for the direction instructions be more tolerant than the generating model. The generating model is a complete, disjunct, partition of the visible area. The acceptor model is a superset of every direction instruction. Therefore, the acceptance areas overlap.

The partitioning into a left and a right half-plane is a sensible acceptor model for the directions "left of" and "right of" relative to the reference object. The dichotomy front/back is modelled similarly by using another axis orthogonal to the reference axis (see figure 2). With some reference frames, however, front and back are exchanged (see below, in the section on three-point localisation). The result is a qualitative distinction, as suggested, for instance, by Freksa [17].

[1] 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.

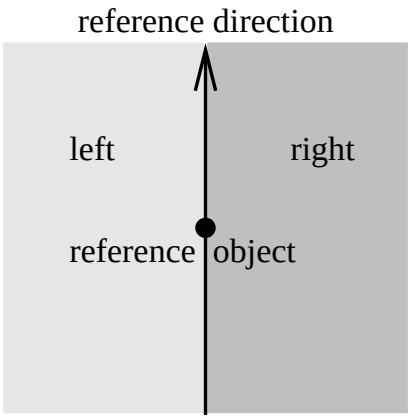


Figure 1. Reference object and reference direction

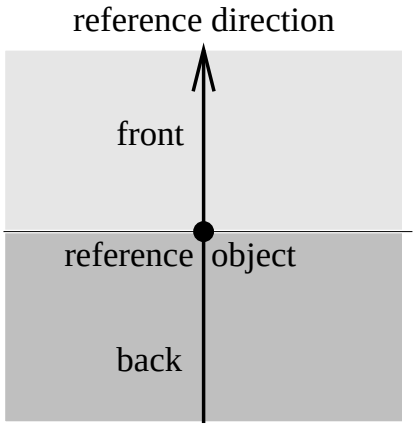


Figure 2. Front-back dichotomy

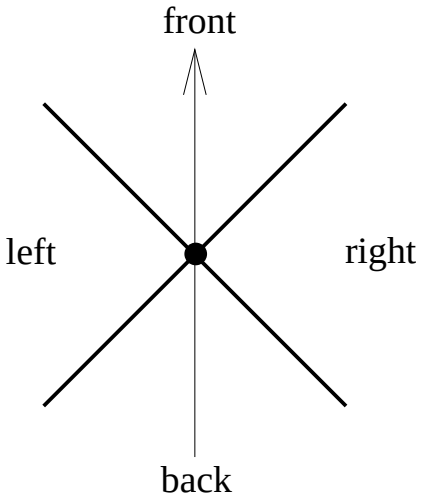


Figure 3. Sector model for generating spatial expressions

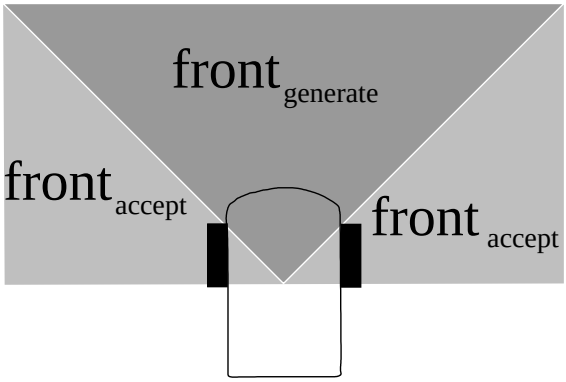


Figure 4. The robot as reference object

To generate spatial expressions, we utilize the simple sector model introduced by Hernández [5]. This model produces four even-sized sectors (see figure 3). The sectors each cover 90 degrees and are prototypical subsets for the corresponding acceptance areas.

#### 2.1.2 Three kinds of reference objects

We need to model spatial references with three kinds of reference objects, namely, the robot, another salient object or a group of objects. If the robot is chosen as reference object, 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. Then, the acceptance area and the generating area for "front" are the ones depicited in figure 4.

If the localisation object is closer to another salient object than to the robot, this object it a convenient reference object. In this case, there are two ways of deriving a reference direction. One is given by the directed straight line from the robot to the the reference object (for instance, through the centers of their projections). This is again adapted to the robot's view. An example for this configuration is shown in figure 5. In this variant of three-point localisation, the "in front of" sector is directed towards the robot. The front/back-dichotomy is inverted, relative to the reference direction [6].

In cases with a group of similar objects, human instructors use references with respect to the whole group, for example, "Fahre zum linken Klotz" ('Go to the left block'). Then the centroid of the group can be treated as the reference object. Analogous to the three point model, the reference direction is given by the directed straight line from the robot center to the the group centroid. This virtual reference object is the origin of acceptance areas and generation areas for relations similar to three-point localisation. The

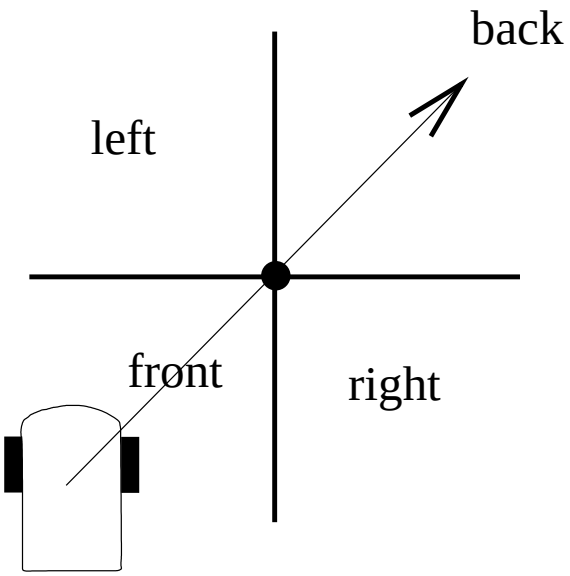


Figure 5. Three point model for generating spatial expressions

object closest to the group centroid can be referred to as the ”middle object”.

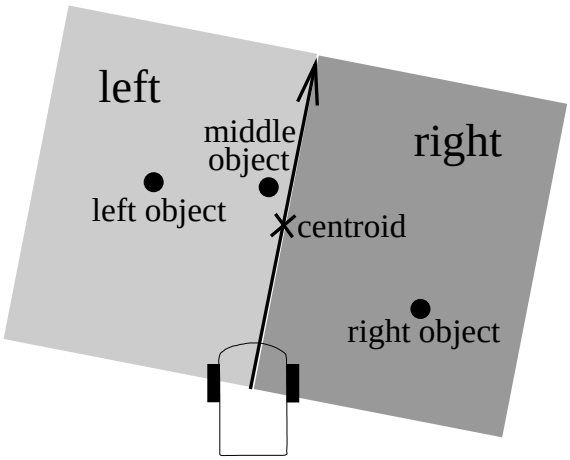


Figure 6. Group based references

We have introduced three kinds of reference systems. The next section describes our natural language robot system. We have experimented with this system in order to determine whether this simple model of spatial references is sufficient to let an instructor pick objects primarily by their location.

Figure 7. Our Robot GIRAFFE.

## 2.2 The Natural Language Controlled Robot System

The architecture of the system is described in [4]. We summarize here the main properties of the system's components. The following components interact in the system: the language interpretation component, the spatial reference component, the sensor component and the acting component.

The **language interpretation component** is based on Combinatory Categorial Grammar (CCG), which has been developed by Steedman [16]. Besides developing a domain dependent version of the lexical categorial system, it was necessary to adapt the grammar rules to German because of the free word order in German and the so-called spurious ambiguity of CCG. Specific categorial rules were derived from the original CCG rules in order to make incremental and efficient processing possible [7].

The **spatial reference component** implements the computational model of projective relations described in section 2.1. It maps the spatial reference expressions of the given command to the relational description delivered from the sensor component. For interpretion commands it uses the acceptor model.

The **sensor component** uses a video camera. An im-

portant decision is to orient to cognitive adequacy in the design of the communicative behavior of the robot, using a sensory equipment resembling human sensorial capabilities [12]. 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 image 2.2). The images are processed with region-based object recognition [12]. The spatial arrangement of these regions is delivered to the spatial reference component as a qualitative relational description.

The **acting component** manages the control of the mobile robot (Pioneer 1). The motoric action the robot can perform are turns and straight movements [15]. The actions can be carried out by passing a control sequence to the motors. The component can do simple obstacle avoidance and path-planning [4].

The interaction between the components consists of a superior instruction-reaction cycle between the language interpretation component and the spatial reference component; subordinate to this cycle is a perception-action cycle started by the acting component, which assumes the planning function and which controls both the sensor component and the acting component.

## 3 Experiments

The questions now asked are:

- 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 spatial reference in human-robot interaction?
- Are the cognitively adequate properties of human-to-human communication cognitively adequate for human-robot interaction? That is, do users use similar strategies for spatial reference in their communication with the robot as they have been found for natural conversation, or do they develop different forms?

### 3.1 Experimental Design

The procedure taken here to answer the above questions was to 'test' the robot in experiments with human users and to see which strategies the users employ initially and how they are adapted during the interaction with the system. Therefore, a test senario was developed in which the user's task was to make the robot move to particular locations pointed at by the leader of the experiment; pointing is used here in order to avoid verbal expression or pictures of the scene which would impose a particular perspective on the scene, for example, the bird's-eye view. Users are instructed to use natural language sentences typed into a computer to move the robot; they are seated in front of a computer in which they type their instructions. When they turn around, they perceive a scene in which, for instance, a number of cubes are placed on the floor together with the robot, which may be placed in a 90 degree angle or opposite of the participant, as shown in figure 8. Thus, depending on the instruction the speaker uses, it becomes clear which point of view she is taking.

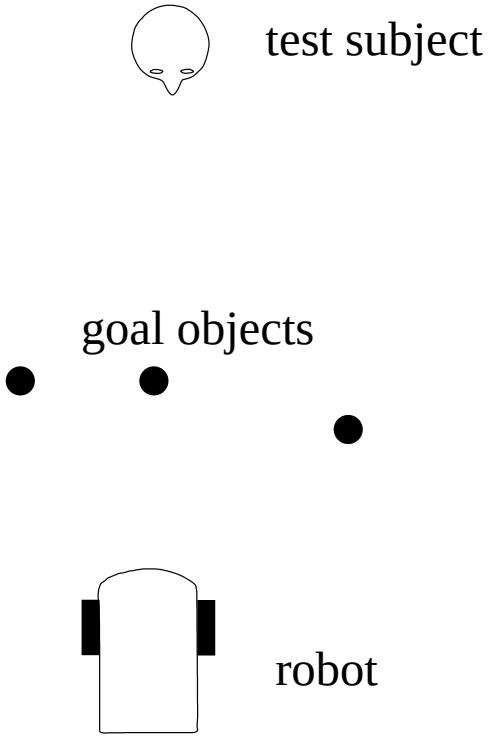


Figure 8. The setting of the experiment

The architecture of the system is described in [4]. 15 different participants carried out about 40 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 anounce their strategies or their ideas about what is going wrong. After the experiments, participants were asked as to what they believe the robot could and could not understand, which strategies they believed to be non-successful, and whether their beliefs about the robot have changed during the interaction. Altogether 603 instructions were elicited.

### 3.2 Experimental Results

Regarding the results gained in research on spatial reference in human-to-human communication reported on in section 2, there are two areas of special interest regarding

the results obtained in the current experiments: the perspective employed and the strategy of instruction.

**Point of View**: While in natural conversation speakers have been found to employ mainly their own point of view, in the experiments the human subjects consistently used the robot's point of view; that they attend to the reference point as a relevant information furthermore becomes clear from two verbal statements recorded during the interaction: In one experiment, the user's first question was where the front of the robot would be, and in another experiment the user had firstly taken the robot's point of view, 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 indeed using her perspective, and she tried out the next instruction accordingly. Sometimes speakers even defined the point of view when they were using instructions that are independent of the perspective taken, for instance: "bewege Dich nach Norden von Dir aus gesehen" 'move north from your point of view'. 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. In human-to-human communication, such a behaviour has been found, for instance, in the interaction with children [11, 6].

**Instructional Strategy**: Regarding object names, participants were found to use mainly basic level categories, such as *cube* or *block*, only one user employing the more abstract category *object*. This is fully consistent with our hypotheses gained on the basis of findings from natural human-to-human communication. However, unlike the strategy consistently used in natural conversation, to name the reference object itself, about half of the participants first tried out much simpler strategies, decomposing the action in more primitive actions, such as *move (a bit) forward*, *go backwards*, or *turn left*, or even *turn your rear wheels*. Half of the participants thus used path specifications or specifications about instrumental, supportative actions to instruct the robot. This strategy seemed very natural to the participants, and they were on the whole quite desparate to find that this strategy did not work with the robot, which, as explained above, was designed on the basis of findings from human-to-human communication. Thus, many of them tried up to 30 sentences before they were hinted at the other strategy by the experimenters, that is, usually they did only reluctantly change their strategy in this respect.

Thus, especially with respect to the perspective taken and the instructional strategies employed, the general ways of how to achieve spatial reference differ considerably in human-robot interaction, compared to human-to-human communication.

## 4 Conclusion

The 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. Therefore, firstly a robot which was implemented as a cognitively adequat model of strategies of spatial reference in human-to-human communication was described. Experiments in human-robot interaction have then been carried out which show that the strategies human users employ to approach an artificial communication partner differ very much from the findings regarding spatial references among humans in two ways: On the one hand, 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. On the other, about half of the participants did not use the strategy from natural communication, to name or describe the goal object, but they described paths how to get there, decomposing the action for the robot. These findings show that designers of human-robot interaction systems cannot solely rely on results from human-to-human communication; furthermore, it seems that what is cognitively adequate modelling depends on the situation which is to be modelled, and that there is no cognitively adequate model of spatial reference which would be independent of the modalities of its employment. Nevertheless, in the design of the current robot, an integration of different modalities, that is, linguistic input, perception and action, was achieved and the future task will be to include also those types of spatial reference which are peculiar to human-robot interaction.

## Acknowledgement

The authors would like to 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 performing the experiments.

## References

[1] 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.

[2] 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 Conference on Human-Computer Interaction, Munich, Germany.*, pages 560–565. Lawrence Erlbaum Ass., London, 1999.

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

[4] 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.

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

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

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

[8] 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.

[9] 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.

[10] G.-A. Levow. Characterizing and recognizing spoken corrections in human-com puter dialogue. In *Proceedings of Coling/ACL '98*, 1998.

[11] 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.

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

[13] S. Oviatt, J. Bernard, and G.-A. Levow. Linguistic adaptations during spoken and multimodal error reso lution. *Langauge and Speech*, 41(3-4):419–442, 1998.

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

[15] R. Röhrig. *Repräsentation und Verarbeitung von qualitativem Orientierungswissen*. Universität Hamburg, Dissertation, Hamburg, 1998.

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

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