Intuitive Linguistic Joint Object Reference in Human-Robot Interaction

Human Spatial Reference Systems and Function-based Categorisation for Symbol Grounding

Reinhard Moratz
Transregional Collaborative Research Center "Spatial Cognition"
Faculty of Mathematics and Informatics, University of Bremen
moratz@informatik.uni-bremen.de

Abstract

The visionary goal of an easy to use service robot implies intuitive styles of interaction between humans and robots. Such natural interaction can only be achieved if means are found to bridge the gap between the forms of object perception and spatial knowledge maintained by such robots, and the forms of language, used by humans, to communicate such knowledge. Part of bridging this gap consists of allowing user and robot to establish joint reference on objects in the environment - without forcing the user to use unnatural means for object reference.

We present an approach to establishing joint object reference which makes use of natural object classification and a computational model of basic intrinsic and relative reference systems. Our object recognition approach assigns natural categories (e.g. "desk", "chair", "table") to new objects based on their functional design. With basic objects within the environment classified, we can then make use of a computational reference model, to process natural projective relations (e.g. "the briefcase to the left of the chair"), allowing users to refer to objects which cannot be classified reliably by the recognition system alone.

Introduction

In a prototypical service robotics task, a human instructor wants the robot to perform an action on a specific object (Zhang & Knoll 2003). Therefore the human and the robot have to establish a joint reference to the objects concerned (Moratz, Fischer, & Tenbrink 2001). In principle, this could be achieved by referring to precise metrical coordinates; or by teaching the robot all potential objects involved, and giving the user a list of proper names (class IDs) of the objects. But then it would be difficult for untrained (naive) users to command the robot; and in a novel situation this simple and unnatural approach would fail. Another strategy would be to use pointing devices, but in 'hand-busy' situations this approach would also fail.

In establishing joint reference, it is essential to avoid the use of forced unnatural communication methods (Thrun 2004); therefore, we present a cognitively inspired approach to establishing joint reference, using only simple, natural linguistic means. In our approach, in order to enable the start of smooth communication between human and robot, the two need to agree on the categorization of a small number of objects in the environment. Even with only this small number of agreed objects available, further communication is then possible, using the initially categorized objects as a frame of reference in the classification of other objects.

For this approach to be successful, it is important to establish a well defined projective relations model to be shared by the robot and the human. For such a model to be useful to naive users, it must be inspired by natural cognitive models, rather than being arbitrarily dictated by the internal workings of the specific robot. Our approach to this referencing uses a well defined qualitative, relative, spatial reference model, which can be used to meaningfully process linguistic terms such as "the bin to the left of the chair".

Object Recognition using Natural Object Categories

Our object recognition module is designed to support unconstrained linguistic access. Human language is built on sets of symbols. The question how these symbols should relate to the physical world is raised within the context of the "Symbol Grounding Problem" (Harnad 1995). According to Harnard (Harnad 1990) a candidate solution to the symbol grounding problem is the use of "categorical representations", which are learned and innate feature-detectors that pick out the invariant features of object and event categories from their sensory projections. In our approach we first want to find out which feature-detectors are powerful enough to detect invariant features of object categories which are important for humans in office scenarios. Therefore we first focus on the design of the feature detectors and only want to add learning capabilities to the system in a second step.

When admitting that objects have certain forms resulting from their functions, this shape can be used to help identify the object's function and ultimately the object itself. We developed an object recognition system handle the input data, render it, and perform our object recognition algorithms [1]. Details of an earlier version of the system have previously been presented in (Wünstel & Moratz 2004).

[1] The system was implemented by my Ph.D. student Michael Wünstel.

In our scenario free-form objects in an office environment have to be identified. In the current version, we mainly focus on the concept of the *supporting plane*. When the function of an object part is to support a potential other object, this part has to be parallel to the ground. A full three-dimensional segmentation based approach is not necessary when additional clues like object arrangement information is given by the user. In the future, we will augment the system with more refined 3D reconstruction abilities. The three-dimensional surface points resulting from the laser range image are separated into up to four layers. We then project each layer into a two-dimensional plane. Within this plane we now can robustly segment object parts by using standard methods. These segments, representing object parts in certain heights, are then used to identify the whole three-dimensional object. The approach performs best for objects having strong functional constraints at the system's current perceptual granularity (e.g. desks, tables, chairs, open doors, and empty book shelfs). However, smaller objects on the ground (e.g. waste paper baskets, briefcases etc.) can be detected but not classified reliably by our current system. These objects can however be referred to by a human, and furthermore they can be referred to by reference to other objects in the environment (e.g. "the briefcase behind the chair"). In the next section we present a model of projective relations which can be used by a robotic agent to facilitate reference to these items which could not otherwise be categorized.

Figure 1 shows an experimental scene together with the scanning equipment. The scene consists of two chairs, a waste paper basket and a briefcase. Figure 2 (b) shows the resulting segments of the lowest level.

Figure 1: Scene together with the scanning equipment on the left

A Computational Model of Projective Relations

An established method in the multidisciplinary research field of Spatial Cognition is to start with psychological findings and theories and build initial computational models of human spatial concepts (Freksa 2004). These computational models can then iteratively be improved based on observations in human-robot interaction experiments.

Figure 2: Resulting segments of the lowest level

Following this method we designed a computational model for projective spatial expressions ("left", "right", "in front of", "behind") based on empirical psychological findings. These spatial location descriptions have been researched extensively in human-human interaction (Herskovits 1986), (van der Zee & Slack 2004).

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 (Levinson 1996) as *intrinsic*, *relative*, and *absolute*.

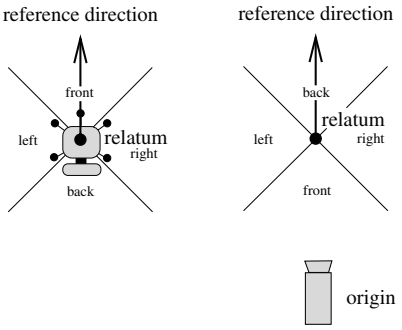
In *intrinsic reference systems*, the relative position of one object (the *referent*) to another (the *relatum*) is described by referring to the relatum's intrinsic properties such as *front* or *back*. In such a situation, the speaker's or hearer's position are irrelevant for the identification of the object. However, the speaker's or hearer's front or back may also serve as origins in intrinsic reference systems.

Humans employing *relative reference systems* use the position of a third entity as origin instead of referring to inbuilt features of the relatum. Thus, a stone (= referent) may be situated to the left of a the house (= relatum) from the speaker's or the hearer's point of view (= origin).

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* serve as anchor directions.

Computational model for linguistic projective relation terms

From such psychological findings (see (Moratz & Tenbrink 2006) for our detailed analysis) we derived a formally defined mapping between linguistic expressions and spatial



(a) Intrinsic ref. model. (b) Relative ref. model.

Figure 3: Acceptance regions (confidence > 0.7) relative to the reference direction.

configurations. Our intention was to keep the corresponding model of projective expressions as simple as possible. This strategy conforms to the widely accepted principle of "Ockham's Razor", which advises that preference be given to the simplest model which explains the data (Popper 1962). Gradually, system refinements can be added which have been proved to be required. Only by starting out with a simple system can the necessary system requirements be worked out. This procedure enables meaningful research in the interdisciplinary and highly complex field of human-robot interaction without the prerequisite of sophisticated and expensive system parts, which may not even be required for enabling effective interaction.

The *computational model* of projective relations implemented in our robot system can be characterised 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 a point-like representation O' of the object O: $O' = \mu(p_{\mathcal{D}}(O))$. For large objects like tables and desks we use an extended model which uses border points of the projected objects (see below).

In intrinsic reference systems the reference direction is given by a direction which is determined by the objects functionality. Figure 3(a) shows a chair's functional direction induced by its back rest. Alternatively, a robot has an intrinsic, functional direction given by its view direction which is by default its front axis.

In relative reference systems, the reference axis is a directed line from the robot center through the relatum, which may be either a group of objects (in which case the group centroid serves as relatum) or (the center of) another salient object in the scenario.

We refer to the angles ϕ_{intr} (for intrinsic reference) and ϕ_{rel} (for relative reference) between the reference direction and the straight line from the relatum to the referent (see

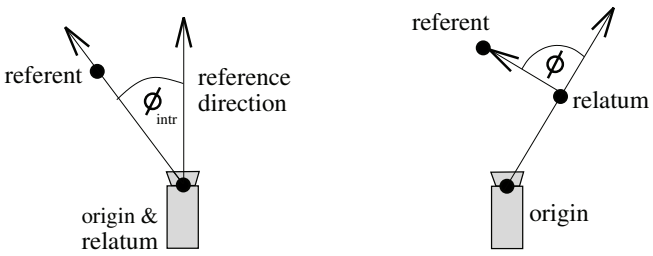


Figure 4: Using ϕ_{intr} and ϕ_{rel} as parameter

figure 4). Each projective term has a *prototypical* direction:

$$
\begin{array}{lcr@{\qquad}lcr}
\phi_{intr}^{\text{Front}} & := & 0^\circ & \phi_{rel}^{\text{Front}} & := & 180^\circ \\
\phi_{intr}^{\text{LeftFront}} & := & 45^\circ & \phi_{rel}^{\text{LeftFront}} & := & 135^\circ \\
\phi_{intr}^{\text{Left}} & := & 90^\circ & \phi_{rel}^{\text{Left}} & := & 90^\circ \\
\phi_{intr}^{\text{LeftBack}} & := & 135^\circ & \phi_{rel}^{\text{LeftBack}} & := & 45^\circ \\
\phi_{intr}^{\text{Behind}} & := & 180^\circ & \phi_{rel}^{\text{Behind}} & := & 0^\circ \\
\phi_{intr}^{\text{RightBack}} & := & -135^\circ & \phi_{rel}^{\text{RightBack}} & := & -45^\circ \\
\phi_{intr}^{\text{Right}} & := & -90^\circ & \phi_{rel}^{\text{Right}} & := & -90^\circ \\
\phi_{intr}^{\text{RightFront}} & := & -45^\circ & \phi_{rel}^{\text{RightFront}} & := & -135^\circ
\end{array}
$$

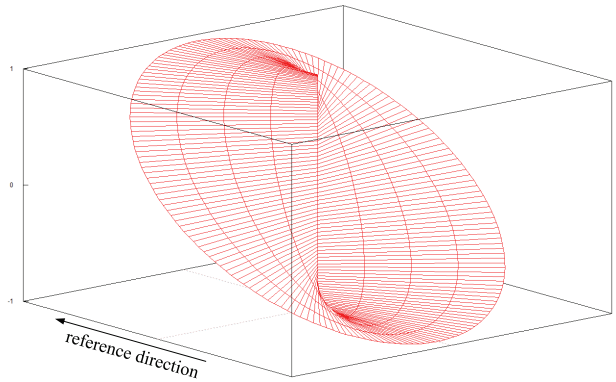


Figure 5: Confidence values for "in front of".

Please note that in relative references the acceptance regions "front" and "back" are exchanged compared to intrinsic references (see figure 3(b)). High confidence values are given to directions which have a smaller angular distance to their corresponding *prototypical* direction (for linguistic findings with respect to the use of prototypes for spatial arrangements see (Herskovits 1986)). We use the term $\cos(\phi - \phi^{\text{prototype}})$ as an expression for our weighting scheme. This weighting scheme generates a smooth ordering function depicted in figure 5 [2].

The linguistic surface of spoken commands does not always reveal the underlying reference system without ambiguities (Moratz & Tenbrink 2006). For example the expression "in front of the chair" can be interpreted ambigously

[2]The assumption of equally treated acceptance regions again conforms to the minimality principle "Ockham's Razor".

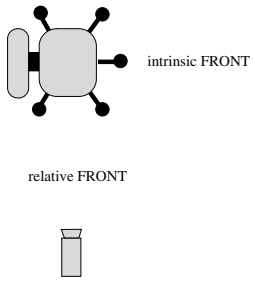
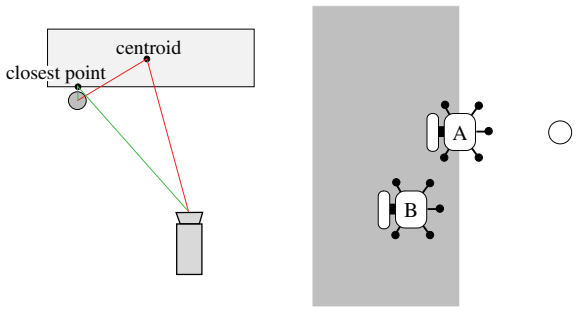


Figure 6: Ambiguous acceptance regions for "in front of the chair".

either as intrinsic (in the direction of the frontal part of the chair) or relative (between the chair and the observer/robot). In specific arrangements the corresponding acceptance regions can be orthogonal (see figure 6) or even diametrically opposed. In order to discover how test subjects comport themselves in these ambiguous situations, we have implemented experiments which will be introduced in the next section.

In cases with a group of similar objects, the centroid of the group (all objects get the same "weight") serves as virtual relatum. The object closest to the group centroid can be referred to as the "middle object". In situations in which a relatum is relative large (e.g. tables, desks) we represent the relatum by the border point of the object's projection which is closest to the referent (see figure 7(a)). Thereby we take the shape of the objects into account. In the configuration depicted in figure 7(a) this difference would lead to the correct linguistic expression "the bin in front of the table" instead of the expression "the bin to the left of the table" which would not be acceptable for human instructors. The results of this feature of our computational model in configurations with simple shaped objects are similar to those of more sophisticated models, which are often inspired by force-field models from physics (Skubic *et al.* 2003) which are needed for more complex object configurations (like walls, bent objects close to each other).



(a) Different angles ϕ for border point (green) versus centroid (red) representation of a table.

(b) Half plane induced by referent and relatum A which shadows the alternative relatum B.

Figure 7: Handling large or multiple relatums.

In standard situations a salient object is only used as relatum if there is no better suited (e.g. closer, more salient) relatum for the referent. We take this into account by defining shadowing regions between a potential referent and pairs of potential relatum objects (see figure 7(b)). The shadowing region is defined as half plane orthogonal to the straight line from the referent to the potential relatum beginning at the relatum.

Using all these mechanisms the system assigns a graded truth value to a potential referent in a spatial configuration with respect to a linguistic expression (maximal truth value for this object after testing all applicable reference systems). Our system takes the object with the highest truth value greater than zero as referent.

In the present stage our system has no dialogue module and therefore cannot handle cases with alternative referents having similar truth values in an interactive manner. In situations where the system cannot solve ambiguities by itself, a more comprehensive system would have to start a clarification subdialogue to resolve these ambiguities.

Experiments and Results

In this section first we look at how our projective relations model was combined with our object recognition system, to facilitate natural, linguistic interaction between human and machine. In our system, users interact by verbally issuing simple requests to the system. These requests - to identify items in the system's perceptual range - are detected with a Nuance Speech Recognizer[3] , before being fed to a semantic analysis component. This analysis attempts to identify the category of the object to be identified, the referent object, and the spatial relationship employed by the user to relate the referent to the target object.

A projection of the recognized 3D objects onto the plane produces a 2D map, defined in terms of object location for directed and undirected objects, object categorization (if available), and camera position and angle. This map is used as input for our integration module; the module - based on the spatial projections model presented in section - gets the spatial knowledge expression/proposition from the semantic analysis component, and attempts to identify the target object in the 2D map using the projective relations defined. The most probable target object, once computed, is then highlighted.

The system's results can be illustrated using an example scene (see figure 1): Figure 8 shows two acceptance regions generated based on the request "show me the briefcase in front of the chair". The two acceptance regions for relative reference are also evaluated. The correct referent (the briefcase, see figures 1 and 8) is selected because it has the highest confidence value. The ranking of the interpretations for our sample configuration is the following. The briefcase in a relative reference system using the left chair as relatum gets the highest confidence value (0.98), the intrinsic interpretation yields a confidence value of 0.81. And the bin between both chairs gets a confidence value of 0.08 for a relative reference interpretation with respect to the chair on the left.

[3] www.nuance.com

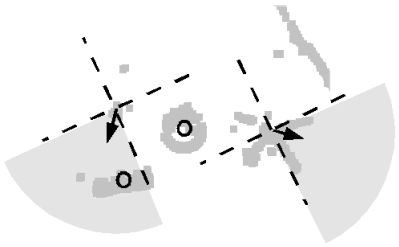


Figure 8: Acceptance areas given a confidence value > 0.7 for the intrinsic interpretation of "in front of the chair".

Now it is a matter of evaluating the spatial reference model of the system in interactions of users. To this end experiments were implemented which will be presented in the next subsections.

Experiments with uninformed users

We designed experiments focussing on configurations in which two possible reference directions can compete (intrinsic/relative to the same relatum). We determined in earlier experiments that the listener/robot perspective will be used with the support of test persons. Now we have, in competition thereto, the relatum direction before directed functional objects which our system recognizes (essentially, the stools in our office scenario).

In conflicting situations the speaker's perspective can naturally be relevant. Our computational model ignores this, which has technically speaking natural advantages [4]. This acceptance of a neglected speaker perspective shall now be particularly tested. Moreover, any preference of an intrinsic or relative reference of test persons shall be tested. In order to be able to systematically vary the position of the speaker/observer *in otherwise identical conditions* it is advantageous to work with a two dimensional abstract imaging simulation (see figure 9 for an image).

In contrast thereto, our earlier efforts (Moratz & Tenbrink 2006), used a real robotic system. The application of a real system naturally represents for a test person a direct motivation, a context, which compares to the goal context of the entire system. The test results thus have an unambiguous validity for real robot systems[5].In order to build the bridge to real systems for the test persons in a standardized way we have this time taken a different path and present the real system to the test persons only in a video. The individual test instructions of the test persons are then based on the abstracted iconified simulation model of the actual system.With this model one can test other languages, test person groups etc. more flexibly.The connection to the real robotic system can be assured in tests using real systems with a very small number of test subjects. Basically, we have determined (Moratz & Tenbrink 2003), that in these settings (application of spatial reference systems without dialogue) one can compare results of speech and typed entries with each other favorably – at worst, there are only small differences.

After presenting the video, the test persons will be invited to give an instruction to a goal object series of configurations "Zeig mir" (which means "Show me") is partly preconfigured: Only a nominal phrase is required to complete the instruction. When the sentence is ready, the test person presses the enter key, the command will be interpreted on the bases of the computational model and the presumed reference object will be highlighted. As a feedback after pressing the button "weiter" (further) either the identified object is highlighted or an error message is presented (either "ich kann den Satz nicht verstehen" which means: "I cannot understand this sentence' e.g. language error, or "ich kann kein passendes Objekt finden" which means: "I cannot find an object that fits" e.g. semantic error).

The test persons are shown an example configuration with an exemplary reference at the end of the film. This applied reference should serve as an example to signal the intended references in order to represent the linguistic means which the parser will process.

To that end one must necessarily apply an intrinsic or relative reference. The example sentence in the film appears only briefly in order to limit any priming as much as possible. In order to estimate/calculate this effect, half of the test persons were given a relative reference as an example sentence, whereas the other half were given an intrinsic reference.

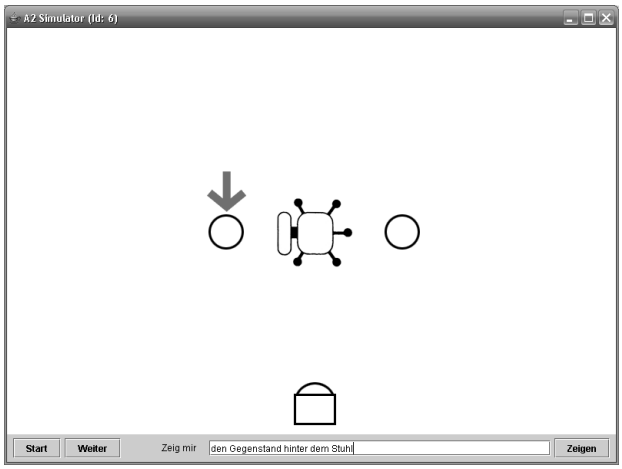


Figure 9: The simulation system.

We designed three different (in various aspects balanced) series of twelve test configurations. We had 18 test subjects, six for each configuration series.

Experimental results

In our experiment from the 216 instructions 165 were successfully parsed by the system. Relative reference was used in 81 configurations, intrinsic reference in 63 configurations.

[4]there are many opportunities for application of the modeling in which the robots can only badly determine the speaker position and/or additional system components were necessary.

[5]They are however also of less validity and reproducability for other research groups and of less transferability to other languages and systems.

In 21 cases the reference failed (e.g. other reference strategies were used, for example counting). We did not find a preference for intrinsic reference in cases when a relative reference made a mental rotation for the test subject necessary (see figure 10).

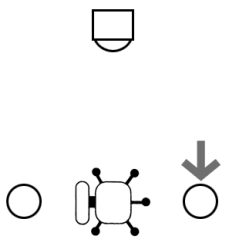


Figure 10: The scanner looks from the opposite direction: A relative reference demands a mental rotation.

Our main observation was that test subjects stick with an at first succesful reference strategy even if this strategy is cognitively more intense. This was shown by the fact that we only observed changes of the reference system with seven of the 18 test subjects. Severeal of these seven test persons which changed the reference system reported after the experiment *curiosity about the system behaviour* as motivation for the changes.

As a result a system designer can build a dialogue strategie which uses the instruction history of the user to infer to the prefered reference system. We had already observed this consequential follow through of a succesful strategy in earlier experiments (Moratz, Fischer, & Tenbrink 2001).

Conclusion and Outlook

The purpose of the system presented in this paper is to offer a simple and natural way for human instructors to inform a robot about the selection of an object even if the category of the object cannot be recognized by the robot. The core of such a system is a mapping function between linguistic expressions and spatial object configurations. To reach this goal a function-based object recognition module, utilizing $2\frac{1}{2}$D laser range data, and a computational model capable of interpreting projective expressions using different reference systems were developed. The distinguishing features of our system are its flexibility and simplicity.

Our model of qualitative spatial arrangements allowed the construction of an interactive reasoning system around the object recognizer, which allowed a user to ask questions concerning the relative positioning of objects recognized within an office environment.

In order for a model of this type to be used by new users without an arduous training phase the reference system model must be cognitively and ergonomically adequate. To that end tests with uninformed users are necessary. We have implemented these type of tests and determined that our computational model of projective relations is in its essential properties adequate.

The full potential of such a system can only unfold if dialogs with the user are possible; Thus, difficulties, particularly involving ambiguities, must be able to be resolved with clarification dialogues. This is the central point of emphasis for further research.

References

Freksa, C. 2004. Spatial cognition - an AI prespective. In *Proceedings of 16th European Conference on AI (ECAI 2004)*.

Harnad, S. 1990. The symbol grounding problem. *Physica D* 42:335–346.

Harnad, S. 1995. Grounding symbolic capacity in robotic capacity. In Steels, L., and Brooks, R., eds., *The Artificial Life route to Artificial Intelligence. Building Situated Embodied Agents*. New Haven: Lawrence Erlbaum. 276–286.

Herskovits, A. 1986. *Language and spatial cognition*. Cambridge, UK: Cambridge University Press.

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

Moratz, R., and Tenbrink, T. 2003. Instruction modes for joint spatial reference between naive users and a mobile robot. In *Proceedings of RISSP IEEE International Conference on Robotics, Intelligent Systems and Signal Processing, Special Session on New Methods in Human Robot Interaction, October 8-13, 2003, Changsha, China*.

Moratz, R., and Tenbrink, T. 2006. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial Cognition and Computation* 6(1):63–107.

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.

Popper, K. 1962. *Conjectures and Refutations.* London: Routledge.

Skubic, M.; Matsakis, P.; Chronis, G.; and Keller, J. 2003. Generating multi-level linguistic spatial descriptions from range sensor readings using the histogram of forces. *Autonomous Robots* 14(1).

Thrun, S. 2004. Toward a framework for human-robot interaction. *Human-Computer Interaction* 19(1-2).

van der Zee, E., and Slack, J. 2004. Representing direction in language and space. *Oxford: Oxford University Press.*

Wünstel, M., and Moratz, R. 2004. Automatic object recognition within an office environment. In *Canadian Conference on Computer and Robot Vision (CRV2004)*.

Zhang, J., and Knoll, A. 2003. A two-arm situated artificial communicator for human-robot cooperative assembly. *IEEE Transactions on Industrial Electronics* 50(4):651–658.