Journal of Universal Computer Science, vol. 9, no. 9 (2003), 1046-1072
submitted: 10/2/03, accepted: 9/6/03, appeared: 28/9/03 © J.UCS

Action Vectors: Modeling Spatial Relations between Objects and Routes

Junko Araki
(University of Tokyo
jun3@is.s.u-tokyo.ac.jp)

Abstract: This paper describes cognitive mechanisms that interpret elliptical instructions used in navigation. We introduce *action vectors*, which are defined as an agent's previous positions on routes. In addition, a new perspective system, *an action-oriented perspective system* is presented. In this system, the action vectors are designated to reference objects. Using theory of the action vectors and the action-oriented system, we demonstrate specific spatial configurations between objects and the action vectors, which arise in cognitive process of interpreting elliptical instructions.

Key Words: navigation system, elliptical route descriptions, perspective systems, human spatial cognition
Category: I.2.4, I.2.7, I.2.10

1 Introduction

This paper presents a unified account of spatial cognition that transforms language into vision. We focus on elliptical route instructions including directional prepositions, which are often used in navigation tasks, and formalize the process of interpreting a series of directional prepositions in sequence during navigation. Regarding the formalization, two problems were still unsolved, one is ambiguity of directional prepositions which has been ignored in cognitive studies, and the other is inadequacy of existing perspective research in order to explain the specific configuration between objects and an agent's previous positions. With respect to the first problem, we identify factor of the ambiguity by introducing *Action Vectors*. The action vectors are defined at the agent's previous positions, where they turned or stopped to interpret each directional preposition (*e.g.*, at an intersection or at an entrance/exit of a room). Introducing action vectors as reference objects demonstrates that the ambiguity of the prepositions is caused by a specific configuration between objects and action vectors. Regarding the second problem, we propose *an action-oriented perspective system*. This system adopts the action vectors as reference objects and demonstrates the configuration between objects and action vectors. Within this research, the directional prepositions are dynamically interpreted taking the continuous interaction of the agent and their environment into account.

The following is a simple example to explain the problem that we deal with in this research. Consider an agent navigating through the world in Figure 1

Table 1: A series of Instructions

1	Go straight down the corridor.
2	Enter the room on the left.
3	Get the document from Tary.
4	Exit the room through the door.
5	Don't forget to turn off the light switch on the left.
6	Go further down the corridor.
7	Enter the third room on the right.
8	Get an apple from Zach.
9	Exit the room through the door.
10	Go straight down the corridor.
11	Enter the room at the end of the corridor.
12	Hand the document and the apple to me.

according to a series of instructions formed as directional prepositions in Table 1. These instructions are provided to the agent before they start navigating. In previous research, these twelve instructions are supposed to be interpreted from the agent's current position on the route in sequence. Thus, the resulting route the agent will follow is shown in Figure 2. Once the agent's view position is put back to their previous positions, another interpretation of these instructions is possible (Figure 3). As an example, consider instruction no.6 ('*Go further down the corridor*'). A second interpretation is considered that results in the emergence of an alternative route in this navigation task. The dotted arrows in Figure 3 show the route resulting from the alternative interpretation. In this case, the agent puts their view position back to the start position and interprets 'go further down' as 'go further down from the agent's start point.' In the same way, the agent interprets instruction no.10 at the view from the exit of Tary's room and arrives at an incorrect room at the opposite end of the corridor. The agent, however, will not find that they are following the incorrect route until the executing task is impossible. The example above implies that the agent refers to their previous positions on the route as reference objects when interpreting route instructions. Unfortunately, no research can explain this interpretation since they do not designate the previous positions as reference objects.

In our theory, we suppose the following situations.

1. A planner adopts elliptic instructions which specify no reference objects (*e.g.*, '*On the left*').

2. An agent interprets a series of instructions in sequence, which is provided as a unified route description in advance, *i.e.,* non-real time navigation.

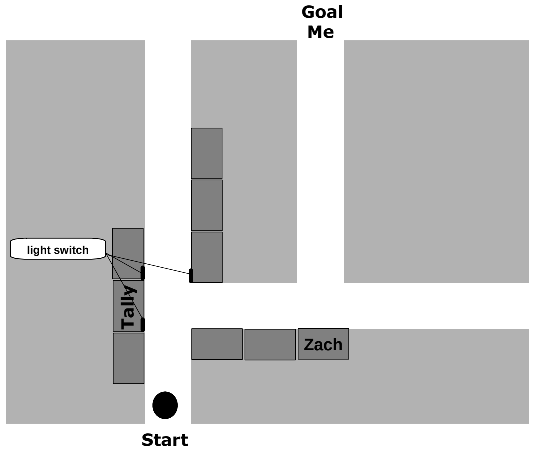

Figure 1: A world where an agent travels according to a series of instructions.

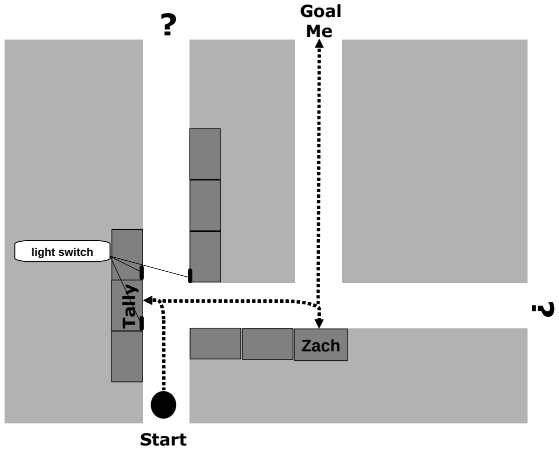

Figure 2: A route assumed by the navigator

2 Related Work

In this section we review main research regarding our study. Our review is comprised of three sections. First, we introduce recent robotic navigation systems which demonstrate their reliability and efficiency even in the face of incomplete information. Secondly, we report classical but developed computational approaches regarding integration of language and vision. In the third, we ex-

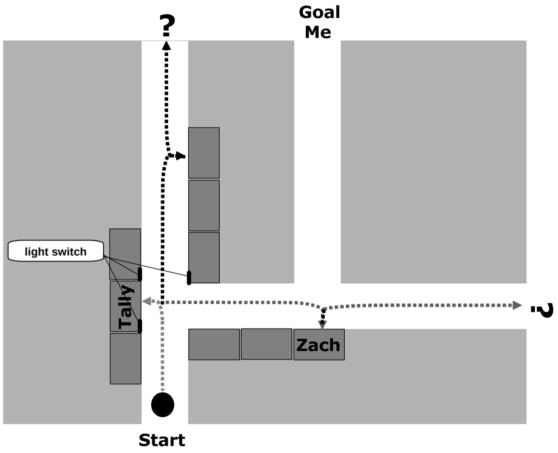

Figure 3: Incorrect routes followed by the agent.

plain existing perspective systems with a figure to lead our main subject in this paper.

2.1 Robotic Approaches to Navigation

Because existing robotic navigation systems have worked by estimating a robot's *current positions* in space, the methods of localization have made remarkable advance during the last decade [Perzanowski et al. 1999, Skubic et al. 2001]. The knowledge of the robot's current positions has been assumed to make it possible for the planner to select the best instruction to the robot and for the robot to head for a destination without getting lost. In other words, all the instructions are considered to be disambiguated only at the robot's current positions on the route. This section introduces successful systems of robotic navigation, which architecture localizes the current positions of the agent.

2.1.1 Simmons

Of several attempts at developing robotic navigation techniques for the decade, Simmons's work [Simmons 1996, Simmons et al. 2002] is both the most concentrated and prolonged research. He introduces an architecture composed of four abstraction layers; navigation, obstacle avoidance, path planning and task scheduling. His research provides much suggestion to ours, especially in concept of his well structured layers: *navigating layer* and *path planning layer*. The layers are discussed below, and demonstrate how his research succeeds in autonomous office navigation.

The **navigation layer** is responsible for getting the robot from one place to another. It uses a Partially Observable Markov Decision Process (POMDP). This model estimates the probability distribution over the positions of the robot at all times, connecting the information about the office environment, approximate distance, and sensor and actuator characteristics together. By robustly tracking the robot's current position, the planner can associate an instruction (*e.g.*, turn or stop) with every Markov state and give the optimal orders to direct the robot to a destination.

The **path planning layer** determines efficient routes based on a topological map augmented with rough metric information and the capabilities of the robot. It uses a decision-theoretic approach to choose the plan with the highest probability of success. For instance, if there is a reasonable chance that the robot will miss seeing a corridor intersection and have to backtrack, a planner might choose a somewhat longer path that avoids that intersection altogether.

Furthermore, these two layers work cooperatively. On the *navigating layer*, a specific Markov model is adopted to compute the robot's current position following the *path planning layer*. As a result, even if the robot might lose its way, the robot moves on another path for a destination.

Other important works of the robotic navigation are techniques of accurate metric precision of the robot's routes. The following traditional methods are widely known: *Configuration Space* by Lozano-Perez [Lozano-Perez 1981], *Generalized Cones* by Brooks [Brooks 1982, Brooks 1986], *Segmented Model* by Crowley [Crowley 1985], *Grid-based Model* by Moravec and Elfes and *Convex Cell model* by Giralt et al [Giralt and Chatila 1979]. We briefly review these methods as follows. Both *Configuration Space* and *Generalized Corns* are developed as solutions to the well-known find path problem in which a robot is to find a collision free, continuous route from their current position. *The Segmented Model* represents the robot's navigational path using a combination of liner segments. The robot's execution path segments have to match the planned segments in both positions and orientation. In *the Grid-based Model*, 2-D horizontal map is adopted to represent the navigable environment. Each cell of the grid cells is probabilistically classified as empty, occupied or uncertain regarding the environment. *The Convex Cell Model* is used in the HILARE system, a mobile robot project at LAAS, France, to represent the world. In this system, free space is represented by connecting the midpoints of adjacent convex cells, which information leads the robot safely to a destination.

2.2 Computational Approaches to Integration of Language and Vision

2.2.1 Gapp

Gapp [Gapp 1994] proposed a multilevel semantic model to compute visual information relating to spatial relations between objects into natural language. His model consists of three levels: geometric level, semantic level and conceptual level. On the geometric level, the geometric properties of objects (basic meanings) are considered according to the theory by Landau, Jackendoff and Langacker [Jackendoff 1993, Landau and Jackendoff 1993, Langacker 1987, Langacker 1991]. This pure information is abstracted on the semantic level and on the conceptual level possible meanings of the spatial relations are represented and their idealized meanings are examined depending on the actual situation and pragmatic factors. In contrast to most existing models, which consider only the 2D case [Zimmer et al. 1998], his model defines semantics of spatial relations in 3D space [Gapp 1995]. An incomplete 3-level system presented by Aurnague [Aurnague and Vieu 1993] inspires the idea of his modularization.

2.2.2 Levitt

Levitt et al. [Levitt and Lawton 1990] propose qualitative methods for navigation in large scale space in their Qualitative Model. The simulated robot uses a global map in its spatial memory to indicate each landmark's estimated direction and distance for route planning. In the topological (qualitative) level, the world is represented as "viewframes" which encodes the observable landmark information that includes angles between landmarks and range estimates. A route is a sequence of locations represented in the qualitative or quantitative map.

2.2.3 Zheng

Zheng et al. [Zeng and Tsuji 1992] presents a unique method in long distance navigation. They use *Panoramic View*, which is a continuous image of sideways view taken from the mobile robot, to represent the navigation route. From the Panoramic View, they extract low-level (*e.g.*, brightness, hue, texture), mid-level (*e.g.,* area, perimeter) and "distinctiveness" of the scene to serve as landmarks. Since this framework assumes that the navigation is executed in a structured environment, such as roads, it is not clear how well the navigation can be adopted for use in an unstructured environment that doesn't provide external guidelines.

2.2.4 The Project VITRA

The project VITRA dealt with the relationship between natural language and vision. Experimental studies had been carried out in the way of designing an interface between image-understanding and natural language systems, with the aim of developing systems for the natural language description of real world image sequences:

In this project, different domains of discourse and communicative situations were examined.

1. Answering questions about observations in a traffic scene [Schirra 1990, Schirra 1992, Schirra and Stopp 1993].
2. Generating running reports for short sections of soccer games [Andre et al. 1989].
3. Describing routes based on a 3-dimentional model of University campus Saarbrucken [Herzog and Wazinski 1994, Maaß 1993].
4. Communicating with an autonomous mobile robot [Leuth et al. 1994].

2.2.5 Maaß

In VITRA project, Maaß constructs a model for the generation of incremental route descriptions[Maaß 1993, Maaß et al. 1995]. An agent named MOSES first takes spatial information and then generates appropriate route descriptions while moving thorough a simulated 3D environment. He criticizes that complete route descriptions generated by cognitive maps do not take the agent's dynamic movements. Thus he adopts the incremental route descriptions, which are generated step by step while MOSES is moving along a path towards a destination. His approach is significant both for more humane navigation systems and for models that transform vision into language in dynamic situations.

2.3 Perspective systems

Perspective systems are cognitive models to determine configurations between objects in space. Depending on both nature of reference objects and the agent's view positions, existing perspective systems are categorized. Countless approaches to the perspective systems exist in the specialized area of psychology and cognitive science [Carlson-Radvansky and Irwin 1993, Carlson-Radvansky and Irwin 1994, Eschenbach and Kulik 1998, Eschenbach and Schill 1999, Herskovits 1986, Langacker 1998a, Langacker 1998b, Levelt 1982, Levelt 1984, Levelt 1989, Miller and Johnson-Laird 1976, Schober 1993, Schober 1995, Schober 1996, Taylor and Tversky 1992, Tversky and Hermenway 1984]. More recently, computational studies apply their approaches to study on integration of vision and natural language processing [Aurnague and Vieu 1993, Gapp 1994, Olivier 1996, Remolina and Kuipers 2002]. Here we review four classical perspective systems closely related to our research as follows.

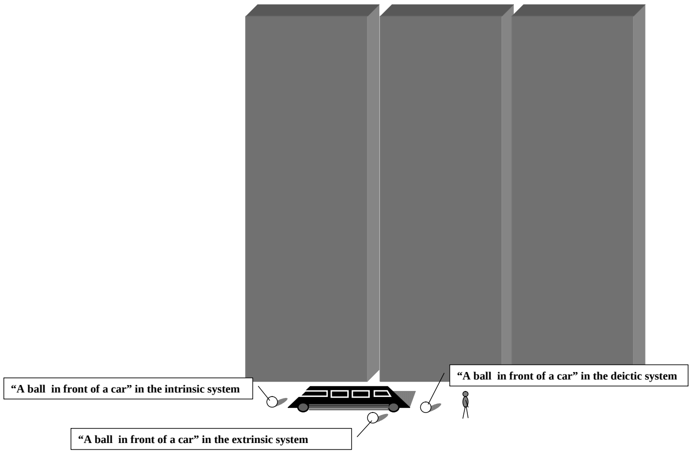
"A ball in front of a car" in the intrinsic system
"A ball in front of a car" in the deictic system
"A ball in front of a car" in the extrinsic system

Figure 4: A world generated a description "A ball in front of a car"

2.3.1 An Intrinsic Perspective System

The intrinsic perspective system is constructed by an orientation of a reference object (*i.e.*, object-centered system). Reference objects like human beings or things which have a kind of faces or front sides are considered their intrinsic orientations to determine directions (intrinsically oriented reference objects). In case where the reference object is the intrinsically oriented object, the spatial configuration between objects is always determined by the intrinsic orientation of the reference object (Figure 4).

2.3.2 A Deictic Perspective System

A deictic perspective system is constructed by an orientation of an agent/viewer (*i.e.*, viewer-centered system). In this system, nature of the reference object (intrinsically oriented or not) are not considered since the spatial configurations is completely determined by the viewer's point of view (Figure 4).

2.3.3 An Extrinsic Perspective System

An extrinsic perspective system adopts an orientation of an *outside* object which exists in close proximity to the reference object. In this system, the outside object is intended to be an intrinsically oriented object and projects its intrinsic orientation on the reference object. Thus, the spatial configuration between objects is secondarily determined by the orientation of the outside object. The figure shows that the front regions of the ball is determined by the building nevertheless the reference object of the ball is not the building but a car (Figure 4). In this case, the intrinsic front of the building is projected on the car, since the building is very near to the car or the viewer has a stronger impact from the building rather than the car. As a result, the inherited front of the car seems to determine the front region of the ball in extrinsic interpretations [Retz-Schmidt 1988].

3 Problems of the Classical Research

We reviewed three kinds of research regarding our study; robotic navigation systems, computational approaches regarding integration of language and vision and the perspective systems. In this section we compare their approaches to ours and point out their problems.

3.1 Problems of Robotic Navigation System

Regarding the navigation systems developing tracking techniques, their weaknesses are the lack of ability to recover from localization failure [Burgard et al. 1998, Skubic et al. 2001]. Most of the tracking techniques maintain and update only a small state space generally centered at the current position of the robot. In other words, they cannot deal with the huge area of a space. Consequently, once they fail in estimating the robot's current position, their navigation stops under complete uncertainty.

Regarding the works in qualitative robotic navigation, they emphasize the importance of landmarks, especially visual landmarks. Few studies, however, provide a basic question "What is a landmark?" The classical research assumes that the landmarks can be readily and easily identified by their internal architectures. Otherwise, they adopt the general definition of "distinctiveness" of features in the sensory readings to identify landmarks. Detailed criteria for defining and selecting plausible landmarks, topological navigation is not possible.

3.2 Problems of Computational Approaches to Spatial Reasoning

The existing computational approaches toward integration of language and vision require accurate metrical information of the distance from the robot's *current* positions to the landmarks or a goal. Thus their architectures are developed to attain precisely estimating or tracking the current position of the agent. As a result, the language descriptions generated by their architecture represent spatial configurations between objects at every static view of the robot. In short, the existing approach doesn't take configuration between the objects and the robot's *previous* views into account [Aurnague and Vieu 1993, Bryant et al. 1992, Gribble et al. 1998, Habel 1990,

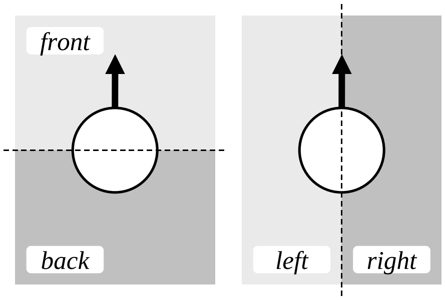

Figure 5: Orientaions of the action vectors

Kray and Blocher 1999, Kray and Porzel 2000, Kray 2001, Kray et al. 2001b, Olivier 1996]. We insist that a dynamic navigation system is a more humane system, where the incremental language descriptions regarding locations of the landmarks or configurations of objects can be produced not only at the robot's current view but at their previous views.

3.3 Problems of the Classical Perspective Systems

As reviewed above, within the classical perspective systems, orientations of existing objects are designated to reference objects. However, an agent moving through space would identify landmarks at their previous positions (*e.g.*, at corners they turned or at an entrance/exit of a building). This specific configuration of an object (a landmark) and the agent previous position is ignored by the classical approaches of perspective study. In order to provide precise explanations to the process of interpreting language descriptions, especially regarding instructions used in navigation, we insist that new dynamic perspective system is desirable.

4 Action Vectors

Action vectors are defined as an agent's previous positions on the route where they turned, stopped or perceived surroundings in a space to execute their missions (*e.g.*, intersections, corners, entrances/exits of rooms). In case where a planner hastily provides elliptical instructions like "Don't forget to turn off the light switch on the left" or "Enter the room on the right" without finding their ambiguous meaning, the agent may designate action vectors (*e.g.*, at a corner they turned or at an entrance/exit of a room) to reference objects and interpret

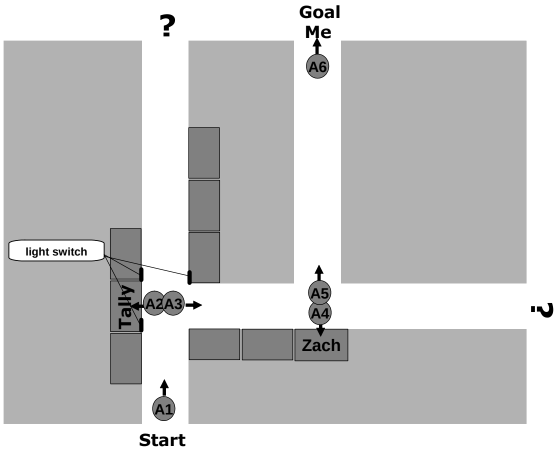

Figure 6: Six action vectors in case of $actvec(c_{10})$

the instructions as “Don’t forget to turn off the light switch on the left of the door” or “Enter the room on the right from view at the corner.” As appeared from this kind of the case, the action vectors can be adopted as reference objects.

Contrasted with other reference objects like trees, the action vectors have intrinsic position vectors and orientation vectors (Figure 5). Let $\mathcal{L}$ ($\subset \mathbf{R}^3$) be the set of all possible position vectors on the map, $\mathcal{O}$ ($\subset \mathbf{R}^3$) be the set of all possible orientation vectors and $\mathcal{C}$ ($= \mathcal{L} \times \mathcal{O}$) be the set of all the possible pairs of a position vector and an orientation vector, which defines the set of action vectors. Then, let D be the description, which is the sequence of the elliptical instructions $d_1 d_2 d_3 \cdots d_n$.

Given the following situations:

- A planner provides a description D to an agent in advance to execute some goal-oriented task.
- At time t, an agent has executed $d_1 d_2 \cdots d_{t-1}$, and is about to execute d_t.

Let c_t be the context that includes the route and any actions the agent had taken at the time t, and *actvec* be the function that takes a context and returns a set of action vectors:

$$\begin{aligned} actvec(c_t) = \{ \; & a_{1,1}, a_{1,2}, \cdots, a_{1,m_1} \\ & a_{2,1}, a_{2,2}, \cdots, a_{2,m_2}, \\ & \cdots, \\ & a_{t,1}, a_{t,2}, \cdots a_{t,m_t} \quad \} \end{aligned}$$

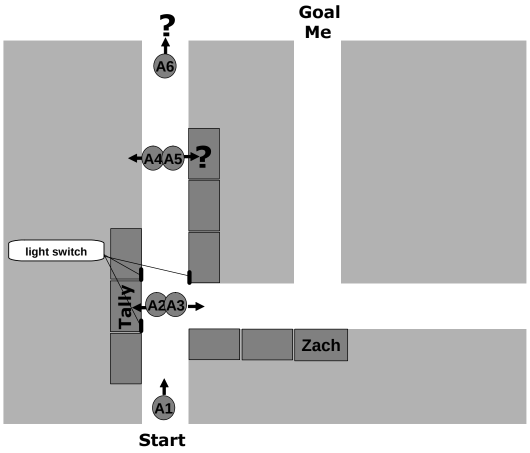

Figure 7: Action vectors generated on the incorrect route

where action vectors $a_{i,1}, \cdots, a_{i,m_i}$ correspond to an instruction d_i, and each action vector $a_{i,j}$ ($\in \mathcal{C}$) is a pair of the position where the agent takes an action according to d_i, and the orientation for which the agent is heading according to d_i. That is, the action vectors temporally and increasingly arise on the route with each instruction. For instance, in case of $actvec(c_{11})$ in the aforementioned navigation task, six action vectors are generated on the route at t_{11} (Figure 6) : the start point (vector a_1) corresponding to the instruction "Go straight down the corridor" (d_1), the entrance of Tary's room (a_2) corresponding to "Enter the second room on the left" (d_2), the exit of Tary's room (a_3) corresponding to "Exit the room through the door" (d_4), the entrance of Zach's room (a_4) corresponding to "Enter the third door on the right" (d_7), the exit of the Zach's room (a_5) to "Exit the room through the door" (d_9), the entrance of my room to "Enter the room at the end of the corridor" (d_{11}).

The case of interest here, the new action vectors generated by the $actvec(c_t)$ become candidates of action vectors to interpret the next instruction at $t+1$. Namely, if the agent chooses a_3 to execute "Go further down the corridor" (d_6), the context of the description will have changed at the time 6 and the action vectors in case of $actvec(c_{11})$, by contrast with the earlier case of $actvec(c_{11})$, will arise on the incorrect route in Figure 7.

4.1 The Action-oriented Perspective System

The principal weakness of the classical perspective systems is the lack of rigor with which our cognitive structure of language comprehension and spatial per-

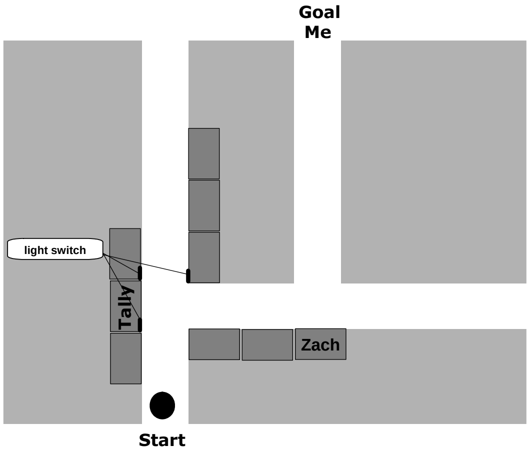

Figure 8: A world where an agent travels according to a series of instructions.

ception is definitely explained [Tversky and Hermenway 1984]. In short, this is due to the failure to add an agent's previous positions (*i.e.*, action vectors) to reference objects. The vast majority of research in the cognitive field misses the point. Thus configurations between objects and the agent's previous positions cannot be systematized.

Contrast to the classical perspective systems, *the action-oriented perspective system* we present here designates the *previous positions* of the agent to the reference objects. The previous positions are already defined as *action vectors* in this section. Adopting action vectors' intrinsic position vectors and orientation vectors, the spatial configurations between the objects and the action vectors are systematically explained. With respect to the specific configurations, we will demonstrate them in the next section.

5 Formalization of Cognitive Mechanisms that Transform Language into Vision

In this section, we formalize cognitive mechanisms that transform language into vision, adopting the action vectors and the action-oriented perspective system. We especially treat with the process of interpreting elliptical instructions. Note that this research does not attempt to disambiguate the elliptical instructions. As long as the instructions are elliptical, the problem of ambiguity does exist. We consistently take an approach of investigating all possible interpretations of the elliptical instructions.

Table 2: A series of Instructions

1	Go straight down the corridor.
2	Enter the room on the left.
3	Get the document from Tary.
4	Exit the room through the door.
5	Don't forget to turn off the light switch on the left.
6	Go further down the corridor.
7	Enter the third room on the right.
8	Get an apple from Zach.
9	Exit the room through the door.
10	Go straight down the corridor.
11	Enter the room at the end of the corridor.
12	Hand the document and the apple to me.

In the aforementioned navigation task (Table 2 and Figure 8), the agent standing at the exit of Tary's room, may interpret "Go into the second room *on the left*" as "Go into the second room *on the left from the start point*"and consequently enter an incorrect room. Still worse the agent, which has successfully arrived at Zach's room, may interpret "*Go further down the corridor*" as "*Go further down the corridor from the exit of Tary's room*" and head for another end of the corridor (Figure 9). Actually Description 1 has three routes derived by different ways of interpretation.

The different interpretations above are caused by the following three factors:

1. the choice of the reference objects

2. the choice of the action vector

3. the choice of the perspective systems

One interpretation for each instruction is determined when these three factors are solved. This process of interpretation is formalized as follows:

Let P be a set of perspective systems, i.e., $P = \{int, dei, aot\}$, where int is the intrinsic perspective system, dei is the deictic perspective system, and aot is the action-oriented perspective system. Given a set of action vectors $actvec(c_t)$ and an instruction d_{t+1}, the agent interprets d_{t+1}. As a result of interpretation, the agent has a *view-frame* v_{t+1} that is defined as $v_{t+1} = \langle p, a, r\rangle$ where p $(\in P)$ is the perspective system that they take, and a $(\in actvec(c_t))$ and r are the action vector and the reference object for the perspective system respectively. The meaning of the view-frame for each perspective system is given as follows: (a) $\langle int, \_, r\rangle$ means an intrinsic system defined for a reference object r, (b) $\langle dei, a, r\rangle$ means a deictic system where r is a reference object and a is a view point, and (c) $\langle aot, a, \_\rangle$ means an action-oriented system defined for an action

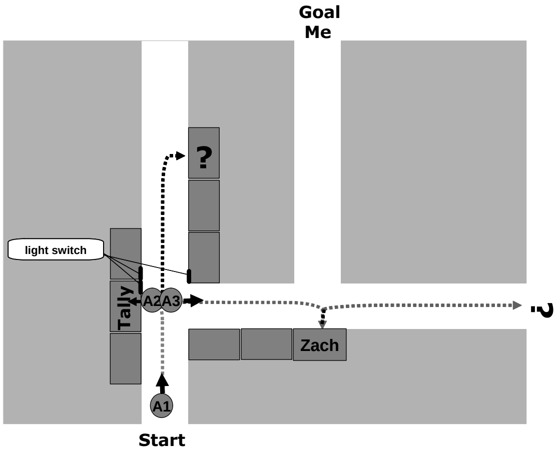

Figure 9: Routes adopting the action vectors as reference objects

vector a.
The finite set of reference objects R is as follows:

$$R = \{\text{ the agent's current positions, action vectors, room doors}\}$$

To demonstrate the concept of our formalization, we describe two examples of the view-frames in Figure 8 as follows:

Example of view-frame (1): The agent navigates according to the instruction $d_1d_2d_3d_4$ in Description 1, and the set of action vectors $actvec(c_4) = \{a_1, a_2, a_3\}$ are defined ((a) in Figure 10). Then the agent turns off the light switch on the left of the door from a_2 (a reference object is the entrance of Tary's room) according to the instruction d_5(=*'Don't forget to turn off the light switch on the left'*). In short, the agent uses the intrinsic left of the door at view from the entrance of Tary's room ((b) in Figure 10). The view frame is constructed as follows.

$$v_5 = \langle int, \_, \text{the door of Tary's room}\rangle$$

The agent, however, may turn off another person's room light switch, which is perceived on the deictic left of the agent standing at the exit of Tary's room (current position) within the deictic system ((c) in Figure 10).

$$v_5 = \langle dei, a_3, \text{the agant's current position}\rangle$$

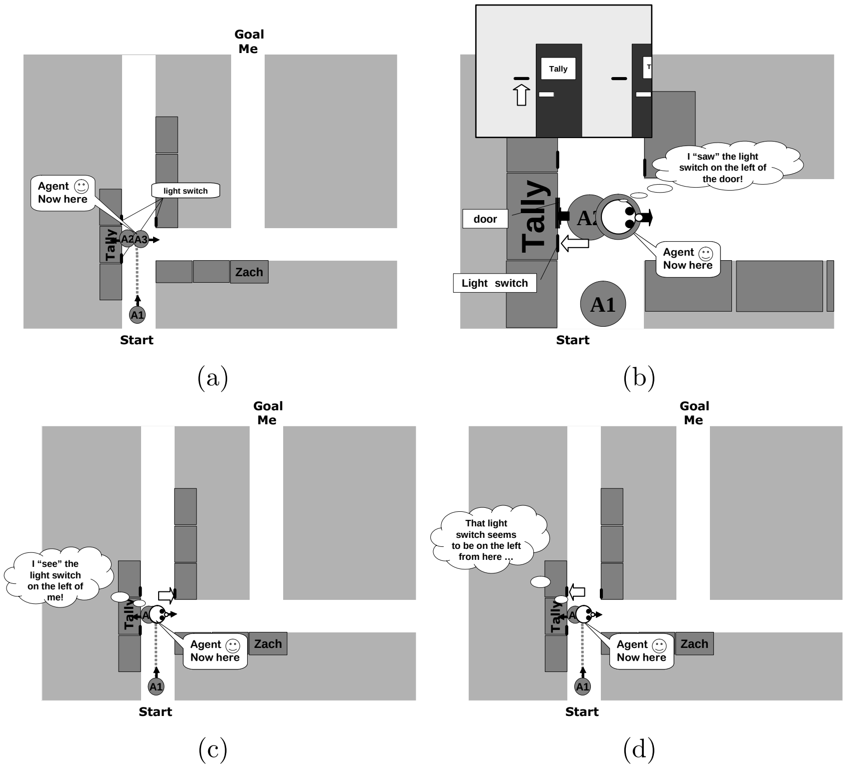

Figure 10: (a) A world of the example view-frame 1, (b) the light switch on the left of the door within the intrinsic perspective system, (c) the light switch on the left of the agent within the deictic perspective system, and (d) the light switch on the left of the agent within the action-oriented perspective system

The agent, however, turns off the light switch on the left of the next door at the view from the action vector a_3 within the action-oriented system ((d) in Figure 10).

$$v_5 = \langle aot, a_3, \_\rangle$$

Example of view-frame (2): The agent navigates in the world according to the instructions $d_1d_2d_3d_4d_5d_6d_7d_8d_9$, and the set of action vectors $actvec(c_9)$ are defined as $\{a_1, a_2, a_3, a_4, a_5\}$ (Figure 11). Then the agent heads for my room (the goal) from the entrance of Zach's room (the current position) according to the instruction d_{10} (=‘*Go straight down the corridor*’). The view-frame is constructed as follows (Figure 12).

$$v_9 = \langle dei, a_5, \text{the agant's current position}\rangle$$

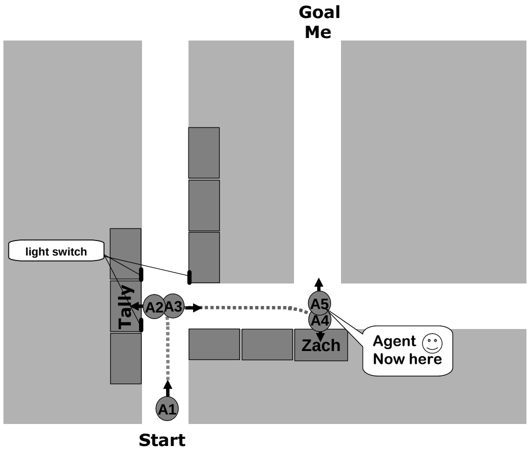

Figure 11: A world of the example view-frame 1

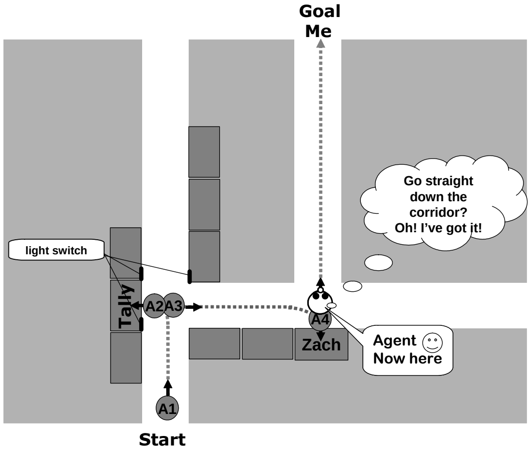

Figure 12: The front direction of the agent within the deictic perspective system

The agent could potentially head for an incorrect room at another end of the corridor from the viewpoint of the action vector a_3 (Figure 13).

$$v_3 = \langle aot, a_3, \_ \rangle$$

In short, cognitive mechanisms that interpret instructions and review the world C is formalized as a sequence of view-frames ($C = v_1 v_2 \cdots v_n$).

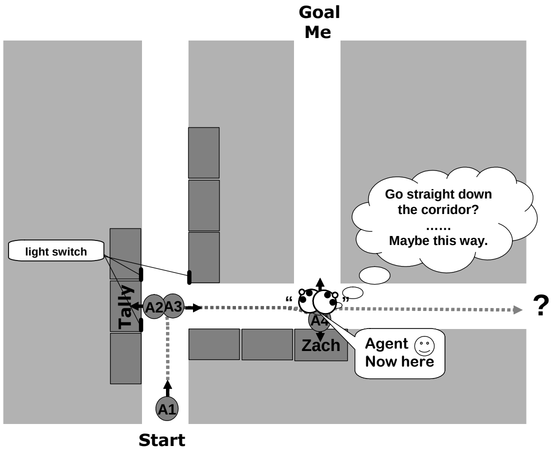

Figure 13: The front direction of the action vector 2 within the action-oriented perspective system

5.1 Incremental Structure of Action Vectors

Action vectors increasingly generate on the route while an agent interprets instructions in sequence. We call the nature of the action vectors *an incremental structure of action vectors*. This structure results from circulation of selecting action vectors and executing instructions by applying adequate view-frames: First, the agent chooses one action vector from existing action vectors (the set of action vectors, *note that the start point is provided as a default*) in order to interpret an instruction. Second the agent moves their next goal according to the intrinsic orientation of the selected action vector. As a result of the agent's move, new action vectors are generated on the route. This circulation is also represented as follows:

With $actvec(c_t)$, the next instruction d_{t+1} and the view-frame v_{t+1}, the next action vectors $a_{t+1,1}, \cdots, a_{t+1,m_{t+1}}$ are determined. Formally, given the instruction d_{t+1} and the view-frame v_{t+1}, we have:

$$actvec(c_{t+1}) = actvec(c_t) \cup f_a(d_{t+1}, v_{t+1})$$

where f_a is a function that corresponds to an action a. f_a takes an instruction and a view-frame and returns the set of newly defined action vectors. Note that in the view-frame v_t, an action vector is selected from the previous set of action vectors $actvec(c_{t-1})$.

For instance, if the agent has got an apple from Zach at t_8 adopting a view-frame v_8, new action vector (a_5) is added to the previous set of action vectors

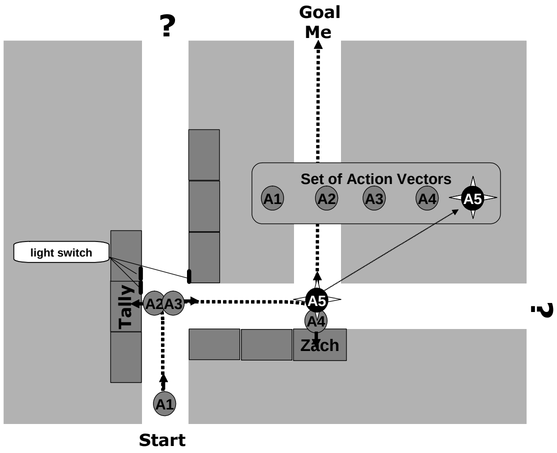

Figure 14: The action vector (a_5) is added to the previous set of action vectors

actvec (c_7) (Figure 14). The formula is as follows;

$$actvec(c_8) = actvec(c_7) \cup f_a(d_8, v_8)$$

6 Application

6.1 Coordination Failures between Subjects in Dot Pattern Descriptions

In this section, we apply our theory of the action vectors and the action-oriented perspective system to interpretations of *dot pattern descriptions*.

Dot pattern descriptions are a kind of instructions to reproduce original configurations between dots in experiments in psychology. In the experiments, subjects are asked to describe instructions in a way that subsequent subjects would be able to reproduce original dot patterns [Jackendoff 1996, Levelt 1989, Levelt 1996]. Yet the reproduced dot patterns are sometimes completely different from the original dot patterns. This problem of *coordination failures* between the first subject and the subsequent subjects is caused by the difference of the perspective systems which the subjects take. In some case, however, the problem of the coordination failures cannot be explained within the theory of the existing perspective systems.

Assume the case where the first subject is given an analogical dot pattern, and is asked to describe an instruction in a way that subsequent subjects would be able to generate it (Figure 15). A typical description by the first subject would

Table 3: A dot pattern description 1

1	Begin with a dot S.
2	Then go straight to dot 1.
3	Draw a right arrow to dot 2.
4	Then draw a left arrow to dot 3.

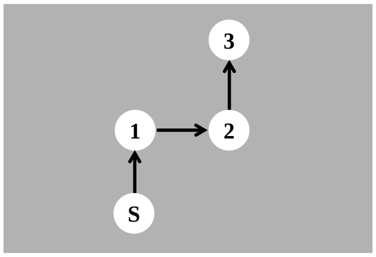

Figure 15: An original dot pattern provided to the first subject.

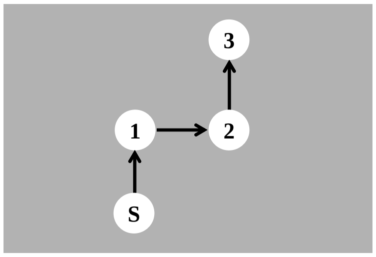

Figure 16: Dot pattern 1 from the dot pattern description 1 adopting the deictic perspective system

be as in Table 3. The description 1, however, would create three dot patterns shown in figure 16, figure 17 and figure 18.

The first subject expresses the description 1 as if they are moving through the diagram or leading the subsequent subjects through it. In short, the first subject adopts the deictic perspective system in the process of transforming the original dot pattern into the description 1 (from *vision* to *language*). Based on the first subject's deictic description, the subsequent subjects attempt to reproduce the same dot pattern as the original one (from *language* to *vision*). Some subsequent subjects also adopt the deictic perspective system and success to reproduce the same dot pattern 1 (Figure 16) as the original one from the description 1. However, other subsequent subjects force the first subject's deictic

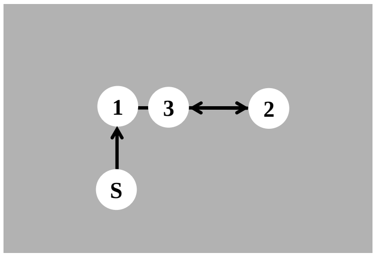

Figure 17: Dot pattern 2 from the dot pattern description 1 adopting the extrinsic perspective system

description into the extrinsic perspective system (Figure 17) and generate dot pattern 2 (Figure 17). In other words, they project their intrinsic orientations on each dot which is flat on the table in front of them, and interpret the third ambiguous sentence "Draw a left arrow to dot 3" of the description 1 adopting the projected orientation of each dot (extrinsic perspective system). As a result, their gaze moves "straight," "right," and "*left*" extrinsically at the last dot they viewed, though the first subject's gaze moves "straight," "right," and "*up*" deictically at the last dot.

This is the simple case of the coordination failures between the first subject and the subsequent subjects generating dot pattern 2. In this case the coordination failures result from the divergence of subjects' perspective into both deictic and extrinsic perspective systems. In short, while the first subject expresses the description 1 in the deictic perspective system, some following subjects interpret the third ambiguous sentence "Draw a left arrow to dot 3" within the extrinsic system. We can easily show the problem using the theory of the classical perspective research as stated above. Problematic case of the coordination failure is in the process of generating dot pattern 3 (Figure 18). As looking at the figure of dot pattern 3 closely, we can find the subsequent subjects' gaze pauses at dot 2 and then *goes back* to dot 1 to execute the third ambiguous sentence "draw a left arrow." The left arrow is actually drawn not from dot 2 but from dot 1. In short, their gaze moves "straight," "right," "left" and "left."

This fact shows that the subsequent subjects don't always interpret the descriptions at **the last dot** they viewed in sequence. They sometimes move their gaze back to some previous and appropriate dot to reproduce the original dot pattern, though their decisions may be wrong. Unfortunately, In the theory of the classical perspective systems, only the last dot (the current dot) is designated to reference objects. In other words, relations of the previous dots and the next dots the subjects will move are not defined in the classical systems. In order to

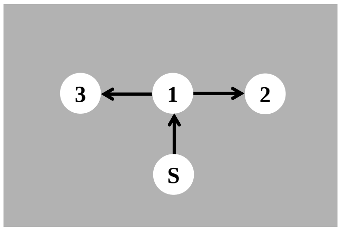

Figure 18: Dot pattern 3 from the dot pattern description 1, which no classical perspective systems can explain

.

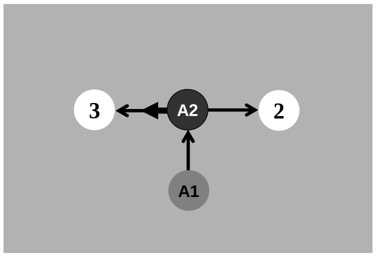

Figure 19: Dot pattern 3 adopting the action vectors and the action-oriented perspective system

explain the process of generating dot pattern 3, our theories of the action vectors and the action-oriented perspective system are inevitably applied as follows.

1. The previous dot S and dot 1 are replaced with *action vectors* $A1$ and $A2$, respectively (Figure 19).
2. $A1$ and $A2$ replaced from dot 1 and dot 2 can determine directions according to intrinsic orientations.
3. When the subsequent subjects execute the third instruction "Draw a left arrow to dot 3," and designate $A2$ rather than $A1$ or the current dot 2 to a reference object, then the left arrow is drawn toward dot 3 according to the intrinsic left of $A2$.

As shown above, the specific coordination failure between subjects arisen in the dot pattern experiment necessitates our theories of the action vectors and the action-oriented perspective system. We contend that our theories can apply to explain more complex cognitive mechanisms that transform language into vision.

7 Conclusion

This research presented new cognitive approach to interpreting route instructions often used in navigation tasks. Based on our mental mechanisms of spatial cognition and language understanding, we examined interpretations of elliptical instructions like "Go straight" or "On the left" and found specific interpretations which were drawn by recognizing objects at view from an agent's previous positions. We systematically demonstrated these specific configurations between the objects and the agent's previous positions by defining *action vectors*. Furthermore we formalized mechanisms of interpreting the elliptical instructions by presenting *an action-oriented perspective system*.

By implementing the above concept, we are developing the computational system, where all the possible meanings of the elliptical instructions are demonstrated and the agent selects one of them through analogy with both the context of route instructions and surroundings.

References

[Andre et al. 1988] Andre, E., Herzog, G., Rist, T.: "On the Simultaneous Interpretation of Real World Image Sequences and Their Natural Language Description: the System SOCCER"; Proceedings ECAI 1988 (1988), 449-454.

[Andre et al. 1989] Andre, E., Herzog, G., Rist,T.: "Natural Language Access to Visual Data: Dealing with Space and Movement. Report 63, Universität des Saarlandes, SFB 314 (VITRA)"; the first Workshop on Logical Semantics of Time, Space and Movement in Natural Language (1989).

[Araki et al. 2002] Araki, J., Ninomiya, T., Makino, T., Tsujii, J.: "Action Vectors for Interpreting Route Descriptions"; Workshop in American Association for Artificial Intelligence (2002).

[Aurnague and Vieu 1993] Aurnague, M., Vieu: "A Three-Level Approach to the Semantics of space"; "The Semantics of Prepositions: from Mental Processing to Natural Language Processing"; Mouton de Gruyter (1993) 393-439, Zelinsky-Wibbelt (editor).

[Blocher and Schirra 1995] Blocher, A., Schirra, J. R. J: "Optional Deep Case Filling and Focus Control with Mental Images: ANTLIMA-KOREF" (1995), 417-423.

[Bryant et al. 1992] Bryant, D., Tversky, B., Franklin, N.: "Internal and External Spatial Frameworks for Representing Described Scenes"; Journal of Memory and Language, 31 (1992), 74-98.

[Brooks 1982] Brooks, R.A.: "Solving the Find-Path Problem by Good Representation Free Space"; Proceedings of AAAI-82 (1982), 381-387.

[Brooks 1986] Brooks, R.A.: "A Robust Layered Control System for a Mobile Robot"; IEEE Journal of Robotics and Automation, RA, 2, 1 (1986) 14-23.

[Burgard et al. 1998] Burgard, W., Derr, A., Fox, D., Cremers, A.: "Integrating Global Position Estimation and Position Tracking for Mobile Robots: the Dynamic Markov Localization Approach"; Proceedings of International Conference on Intelligent Robots and Systems (IROS) (1998).

[Carlson-Radvansky and Irwin 1993] Carlson-Radvansky, L.A., Irwin, D.: "Frames of Reference in Language and Vision: Where is above?"; Cognition, 46 (1993), 223-244.

[Carlson-Radvansky and Irwin 1994] Carlson-Radvansky, L.A., Irwin, D.: "Reference Frame Activation during Spatial Term Assignment"; Journal of Memory and Language, 33 (1994), 646-671.

[Crowley 1985] Crowley, J.L.: "Navigation for an Intelligent Mobile Robot"; IEEE Journal of Robotics and Automation, RA, 1, 1 (1985) 31-41.

[Eschenbach and Kulik 1998] Eschenbach, C., Kulik, L.: "An Axiomatic Approach to the Spatial Relations Underlying Left Right and in Front of-Behind"; KI-97 Advances in Artificial Intelligence (1998).

[Eschenbach and Schill 1999] Eschenbach, C., Schill, K.: "Studying Spatial Cognition: A report on the DFG workshop on "The Representation of Motion" "; Kunstliche Intelligenz, 13, 3 (1999), 57-58.

[Freeman 1975] Freeman: "The Modeling of Spatial Relations"; Computer Graphics & Image Processing, 4 (1975), 156-171.

[Gapp 1994] Gapp, K. P.: "Basic Meanings of Spatial Relations: Computation and Evaluation in 3D Space"; AAAI-94 (1994), 95-105.

[Gapp 1995] Gapp, K.P.: "Angle, Distance, Shape, and their Relationship to Projective Relations"; Proceedings of the 17th Conference of the Cognitive Science Society (1995).

[Giralt and Chatila 1979] Giralt, G., Chatila, R.: "A multi-level planning and navigation system for a mobile robot"; Proceedings of IJCAI-79 (1979) 335-337.

[Gribble et al. 1998] Gribble, W. S., Browning, R. L., Hewett, M., Remolina, E., Kuipers, B. J.: "Integrating Vision and Spatial Reasoning for Assistive Navigation"; Lecture Notes in Computer Science, 1458 (1998).

[Habel 1990] Habel, C.: "Propositional and Depictorial Representations of Spatial Knowledge: The Case of Path-concepts"; Springer Verlag (1990).

[Herskovits 1986] Herskovits, A.: "Language and Spatial Cognition. An Interdisciplinary Study of the Prepositions in English"; Cambridge University Press (1989).

[Herzog and Wazinski 1994] Herzog, G., Wazinski, P.: "VIsual TRAnslator: Linking Perceptions and Natural Language Descriptions"; Artificial Intelligence Review, 8, 2-3 (1994), 175-187.

[Herzog and Rohr 1995] Herzog, G., Rohr, K.: "Integrating Vision and Language: Towards Automatic Description of Human Movements"; Kunstliche Intelligenz (1995), 257-268.

[Jackendoff 1993] Jackendoff, R.S.: "Semantic Structures"; MIT Press (1993).

[Jackendoff 1996] Jackendoff, R.: "Language and Space", chapter "The Architecture of the Linguistic-Spatial Interface"; Bloom, P., Peterson, M.A., Garrett, M.F., Nadels, L. (eds.), The MIT Press (1996), 1-30.

[Jorrand and Sgurev 1987] Jorrand, K., Sgurev, L.: "Artificial Intelligence II: Methodology, Systems, Applications" (1987) 375-382.

[Kray and Blocher 1999] Kray, C., Blocher, A.: "Modeling the Basic Meanings of Path Relations"; IJCAI (1999), 384-393.

[Kray and Porzel 2000] Kray, C., Porzel, R.: "Spatial Cognition and Natural Language Interfaces in Mobile Personal Assistants"; ECAI Workshop on Artificial Intelligence in Mobile Systems, AIMS-2000, IJCAI (2000).

[Kray 2001] Kray, C.: "The Benefits of Multi-Agent Systems in Spatial Reasoning"; Proceedings of Flairs'01 (2001), 552-556.

[Kray et al. 2001a] Kray, C., Baus, J., Zimmer, H., Speiser, H., Krüger, A.: "Two Path Prepositions: Along and Past"; Lecture Notes in Computer Science, 2205 (2001).

[Kray et al. 2001b] Kray, C., Baus, J., Zimmer, H., Speiser, H., Kruger, A.: "Two Path Prepositions: Along and Past"; Proceedings of COSIT'01 (2001).

[Landau and Jackendoff 1993] Landau, B., Jackendoff, R.: ""What" and "where" in Spatial Language and Spatial Cognition"; Behavioral and Brain Sciences, 16, 2 (1984), 217-238.

[Langacker 1987] Langacker, R.W.: "Foundations of Cognitive Grammer: Theoretical Prerequisites Vol. 1"; Stanford University Press (1987).

[Langacker 1991] Langacker, R.W.: "Foundations of Cognitive Grammer: Descriptive Application Vol. 2"; Stanford University Press (1991).

[Langacker 1998a] Langacker, R.W.: “Topics in Cognitive Linguistics”, chapter “An Overview of Cognitive Grammar”; Rudzka-Ostyn, B. (eds.), John Benjamins Publishing Company (1998), 3-48.

[Langacker 1998b] Langacker, R.W.: “Topics in Cognitive Linguistics”, chapter “A View of Linguistic Semantics”; Rudzka-Ostyn, B. (eds.), John Benjamins Publishing Company (1998), 49-90.

[Leuth et al. 1994] Lueth, T., Laengle, T., Herzog, G., Stopp, E., Rembold, U.: “Kantra: Human-Machine Interaction for Intelligent Robots Using Natural Language”; Proceedings of IEEE International Workshop on Robot and Human Communications (1994), 106-111.

[Levelt 1982] Levelt, W.J.M.: “Speech, place, and action: Studies in deixis and related topics”, chapter “Cognitive Styles in the Use of Spatial Direction Terms”; Jarvella, R.J., Klein, W. (eds.), Wiley (1982), 251-268.

[Levelt 1984] Levelt, W.J.M.: “Limits in perception”, chapter “Some Perceptual Limitations on Talking about Space”; van Doorn, A.J., van de Grind, W.A., Koenderink, J.J. (eds.), VNU Science Press (1984).

[Levelt 1989] Levelt, W.J.M.: “Speaking: From Intention to Articulation”, MIT Press (1989).

[Levelt 1996] Levelt, W.J.M.: “Language and Space”, chapter “Perspective Taking and Ellipsis in Spatial Descriptions”; Bloom, P., Peterson, M.A., Garrett, M.F., Nadels, L. (eds.) The MIT Press (1996), 77-107.

[Levitt and Lawton 1990] Levitt, T.S., Lawton, D.T.: “Qualitative Navigation for Mobile Robots”; Artificial Intelligence (1990), 305-360.

[Lozano-Perez 1981] Lozano-Perez: “Automatic Planning of Manipulator Transfer Movements”; Proceedings of IEEE Transactions on Systems Mans and Cybernetics, SMC (1981), 681-698.

[Maaß 1993] Maaß, W.: “A Cognitive Model for the Process of Multimodal, Incremental Route Description”; Proceedings of the European Conference on Spatial Information Theory (1993).

[Maaß 1994] Maaß, W.: “From Vision to Multimodal Communication: Incremental Route Descriptions”; Artificial Intelligence Review, 8, 2-3 (1994), 159-174.

[Maaß et al. 1995] Maaß, W., Baus, J., Paul, J.: “Visual Grounding of Route Descriptions in Dynamic Environments”; Computational Models for Integrating Language and Vision, AAAI 1995 Fall Symposium Series (1995).

[Miller and Johnson-Laird 1976] Miller, G.A., Johnson-Laird, P.N.: “Language and Perception”; Harvard University Press (1976).

[Olivier 1996] Olivier, P.: “Mediating the Interchange of Information between the Visual and Verbal Domains”; Ph. D. Thesis, The University of Manchester Institute of Science and Technology (1996).

[Perzanowski et al. 1999] Perzanowski, D., Schultz, A., Marsh, E., Adams, W.: “Goal Tracking in a Natural Language Interface: Towards Achieving Adju stable Autonomy”; Proceedings of the 1999 IEEE International Symposium on Computational Intelligence in Robotics and Automation (1999), 144-149.

[Pick and Acredolo 1983] Pick, H., Acredolo, L.: “How language structures space. Spatial Orientation: Theory, Research and Application”; Plenum Publishing Corp (1983).

[Remolina and Kuipers 2002] Remolina, E., Kuipers, B.: “Towards a General Theory of Topological Maps”; Submitted to Artificial Intelligence Journal (2002), 24-25.

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

[Schirra 1990] Schirra, J. R. J.: “A Contribution to Reference Semantics of Spatial Prepositions: The Visualization Problem and its Solution in VITRA”; Bericht Nr. 75, Informatik, Universität des Saarlandes, Saarbrücken, Digital Equipment Corporation, Saarbrücken (1990).

[Schirra 1992] Schirra, J. R.: "Connecting Visual and Verbal Space: Preliminary Considerations Concerning the Concept 'Mental Image' ";Proceedings of the Fourth European Workshop on Semantics of Time, Space and Movement and Spatio-Temporal Reasoning (1992).

[Schirra and Stopp 1993] Schirra J. R. J., Stopp, E.: "ANTLIMA - A Listener Model with Mental Images"; IJCAI (1993), 175-180.

[Schober 1995] Schober, M.: "Speakers, Addressees and frames of reference: Whose Effort is Minimized in Conversations about Locations?"; Discourse Processes, 20 (1995), 219-247.

[Schober 1993] Schober, M.: "Spatial Perspective-Taking in Conversation"; Cognition, 47 (1993), 219-247.

[Schober 1996] Schober, M.: "Addressee- and Object-Centered Frames of Reference in Spatial Descriptions"; American Association for Artificial Intelligence, Working Notes of the 1996 AAAI Spring Symposium on Cognitive and Computational Models of Spatial Representation, 47 (1996), 92-100.

[Simmons and Koenig 1995] Simmons, R., Koenig, S.: "Probabilistic Robot Navigation in Partially Observable Environments"; Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95) (1995), 1080-1087.

[Simmons 1996] Simmons, R.: "The Curvature-Velocity Method for Local Obstacle Avoidance"; International Conference on Robotics and Automation (1996).

[Simmons et al. 1996] Simmons, R., Henricksen, L., Chrisman, L., Whelan, G.: "Obstacle Avoidance and Safeguarding for a Lunar Rover"; Proceedings AIAA Forum on Advanced Developments in Space Robotics (1996).

[Simmons et al. 1997a] Simmons, R., Koenig, S., Lopez, J., Goodwin, R.: "Towards Self-Reliant Autonomous Systems"; Workshop on Planning and Scheduling for Space (1997).

[Simmons et al. 1997b] Simmons, R., Goodwin, R., Zita, H. K., Koenig, S., O'Sullivan, J.: "A Layered Architecture for Office Delivery Robots"; First International Conference on Autonomous Agents (1997).

[Simmons and Thrun 1998] Simmons, R., Thrun, S.: "Languages and Tools for Task-Level Robotics Integration"; AAAI Spring Symposium on Integrating Robotics Research (1998).

[Simmons and Apfelbaum 1998] Simmons, R., Apfelbaum, D.: "A Task Description Language for Robot Control"; Proceedings of Conference on Intelligent Robotics and Systems (1998).

[Simmons and Pecheur 2000] Simmons, R., Pecheur, C.: "Automating Model Checking for Autonomous Systems"; Proceedings of the AAAI Spring Symposium on Real-Time Autonomous Systems (2000).

[Simmons et al. 2000a] Simmons, R., Apfelbaum, D., Burgard, W., Fox, D., Moors, M., Thrun, S., Younes, H.: "Coordination for Multi-Robot Exploration and Mapping"; Proceedings National Conference on Artificial Intelligence (2000).

[Simmons et al. 2000b] Simmons, R., Pecheur, C., Srinivasan, G.: "Towards Formal Verification of Autonomous Systems"; In Proceedings of the Conference on Intelligent Robots and Systems (IROS) (2000).

[Simmons et al. 2002] Simmons, R., Smith, T., Dias, M. B.,Goldberg, D., Hershberger, D., Stentz, A., Zlot, R.: "Multi-Robot Systems: From Swarms to Intelligent Automata", chapter "A Layered Architecture for Coordination of Mobile Robots"; Schultz, A., Parker, L. (eds.), Kluwer (2002).

[Skubic et al. 2001] Skubic, M., Matsakis, P., Forrester, B., Chronis, G.: "Probabilistic Robot Navigation in Partially Observable Environments"; Proceedings of the 2001 IEEE International Conference on Robotics and Automation (2001).

[Skubic et al. 2002] Skubic, M., Matsakis, P., Chronis, G., Keller, J.: "Generating Multi-level Linguistic Spatial Descriptions from Range Sensor Readings Using the Histogram of Forces"; Preparation for submission to Autonomous Robots (2002).

[Taylor and Tversky 1992] Taylor, H., Tversky, B.: “Descriptions and depictions of environments”; Memory and Cognition, 20 (1992), 483-496.

[Tversky and Hermenway 1984] Tversky, B., Hermenway, K.: “Objects, parts and categories”; Journal of Experimental Psychology: General, 113, 2 (1984), 169-191.

[Zeng and Tsuji 1992] Zeng, J.Y., Tsuji, S.: “Panoramic Representation for Route Recognition by a Mobile Robot”; International Journal of Computer Vision, 9, 1 (1992), 55-76.

[Zimmer et al. 1998] Zimmer, H.D., Speiser, H.R., Baus, J., Blocher, A., Stopp, E.: “The Use of Locative Expressions in Dependence of the Spatial Relation Between Target and Reference Object in Two-Dimensional Layouts”; Lecture Notes in Computer Science, 1404 (1998).