Sonderforschungsbereich 314
Künstliche Intelligenz - Wissensbasierte Systeme

KI-Labor am Lehrstuhl für Informatik IV

Leitung: Prof. Dr. W. Wahlster

Universität des Saarlandes
FB 14 Informatik IV
Postfach 151150
D-66041 Saarbrücken
Fed. Rep. of Germany

Tel. 0681 / 302-2363

VITRA

Bericht Nr. 116

How Spatial Information Connects Visual Perception and Natural Language Generation in Dynamic Environments: Towards a Computational Model

Wolfgang Maaß

Juni 1995

ISSN 0944-7814

116

How Spatial Information Connects Visual Perception and Natural Language Generation in Dynamic Environments: Towards a Computational Model

Wolfgang Maaß

Department for Computer Science
Universität des Saarlandes
Im Stadtwald 15
D-66041 Saarbrücken 11, Germany

E-Mail : maass@cs.uni-sb.de
Phone: (+49 681) 302-3393
Fax: (+49 681) 302-4421

Abstract. Suppose that you are required to describe a route step-by-step to somebody who does not know the environment. A major question in this context is what kind of spatial information must be integrated in a route description. This task generally refers to two cognitive abilities: Visual perception and natural language. In this domain, a computational model for the generation of incremental route descriptions is presented. Central to this model is a distinction into a visual, a linguistic, and a conceptual-spatial level. Basing on these different levels a software agent, called MOSES, is introduced who moves through a simulated 3D environment from a starting-point to a destination. He selects visuo-spatial information and generates appropriate route descriptions. It is shown how MOSES adopts his linguistic behavior to spatial and temporal constraints. The generation process is based on a corpus of incremental route descriptions which were collected by field experiments. The agent and the 3D environment are entirely implemented.

1 Introduction

What kind of spatial information is necessary for the provision of incremental route descriptions? This question combines two important cognitive modules: Visual perception and natural language generation. We present a computational model of a situated agent[1], called MOSES. He[2] moves through an unknown simulated urban-like 3D environment (see figure 2). His task is to select a path from a map and to describe appropriate actions step-by-step to a virtual listener moving along this path. In contrast to comparable models, MOSES does not simply access information about the environment from a database. MOSES has rather a visual perception module which allows him to perceive and select information from the simulated environment. This approach is grounded on research results gained during a cooperation with the visual perception group of the IIFB at the Fraunhofer Institute, University of Karlsruhe. In joint research with this group we investigated how a model-based approach for visual object selection can be used to automatically recognize 3D object representations. In several domains, we examined how real world data can be used in natural language description systems, e.g. in a soccer domain (cf. [André et al. 89]) and in a traffic domain (cf. [Schirra et al. 87]). The model presented here is based on these investigations.

Studies in visual perception, such as Marr's influential work (cf. [Marr 82]), investigate how visual information is used to construct an internal spatial representation of visible objects. Marr did not, however, show any links between 3D model representations and other processes using spatial information such as natural language processing. Researchers working in this area have mainly been con-

[1] We define a situated agent as a computational module which acts in virtual or real environments. It consists of one or more decision making modules and knowledge bases. Its reasoning and planning abilities mainly depend on perception and on self-obtained knowledge. Therefore its knowledge is generally incomplete and inconsistent. But a situated agent is able to adapt its behavior to changes in given situations of the environment.

[2] For readability reasons we use male forms while referring to MOSES throughout this article

cerned with problems related to the recognition of single objects and object parts (e.g., [Marr & Nishihara 78; Binford 71]). The main purpose of visual perception is to select and group information units in order to make sense out of basic sensor stimulations. Although selection of information is a major issue, most systems only investigate this at early levels of visual processing. Whenever we perceive our environment we select information. The process of visual selection presumes that information provided by the environment is much richer and more complex than what a perceptual system is able to process. Hence, it is assumed that our cognitive system constructs a spatial mental model of the current environment. From this perspective, visual perception is important as input for independent conceptual-spatial representations (e.g., [Johnson-Laird 83]). On the other hand, approaches from linguistics consider the linguistic structure used for describing configurations as being fundamental for spatial cognition (e.g., [Talmy 83; Lakoff 87; Herskovits 86]). Herskovits, for instance, recognizes the distinction between a spatial level and a linguistic level of spatial terms but she does not investigate the relation between both levels in detail (cf. [Herskovits 86, p.102]).

What is known though about a mode-independent spatial level? Baddeley and Hitch proposed in their working memory theory an independent module, called the *visuo-spatial sketch pad* (cf. [Baddeley & Hitch 74; Baddeley 86]). It is concerned with the temporary storage of visuo-spatial information. Another advocator who adopts a linguistic perspective is Jackendoff. Although his *conceptual structure* is strongly influenced by linguistic considerations he writes: "There is a *single* level of mental representation, *conceptual structure*, at which linguistic, sensory, and motor information are compatible"[3] ([Jackendoff 83, p.17]). Another approach is proposed by Johnson-Laird, who states that Marr's general assumption that "all our knowledge of the world depends on our ability to construct models of it" is the basics for all computational models of cognitive processes ([Johnson-Laird 83, p.402]). By his *mental model* approach, Johnson-Laird also

[3] Italics are from the original text.

suggests an independent knowledge structure between cognitive modules. As Johnson-Laird points out, "we have no way of knowing what the structure is (or even whether the notion makes sense) that is independent from the way in which we conceive the world" ([Johnson-Laird 83, p. 402]). Based on ideas of mental models, Bryant outlined a spatial representation system (SRS) in which he stressed the importance of differernt kinds of frames of reference (cf. [Bryant 92]). Couclelis presents in her proposal how pre-conceptual schema representations, mental models, and cognitive maps can be seen as based on one another (cf. [Couclelis 94]).

A fundamental question for a complete computational theory dealing with the integration of natural language and visual perception is what kind of processes and representations lie in-between[4]. It is fairly well established that visual perception and natural language are independent, cognitive modules and that they have their own representations and processes. An implicit assumption for combining both systems is to look for well-suited interfaces. Similar to retinotopic projections, visual information obtained from a given situation first of all provides two-dimensional information projected onto a plane orthogonal to the direction of movement (see figure 1 which illustrates a crossing scenario). It can be directly distinguished between those objects on the left, those on the right, and those in front. The same holds for top and bottom. More complex to obtain is information about how items are ordered relative to one another. For instance, the relation that item A is behind item B generally requires commonsense knowledge about these items as well as stereo-vision.

Central to the model proposed here is a distinction between mode-specific and mode-independent representations of spatial knowledge. Representations associated to visual perception and natural language are mode-dependent. Conceptual and in particular spatial information is assumed to be processed and represented

[4] The integration of visual processing and natural language processing is currently a new and hotly discussed topic in AI (cf. [McKevitt 94b; McKevitt 94a]).

at a mode-independent level in-between. Representations and processes at this level are not understood in detail. In experiments data is almost exclusively obtained by verbal descriptions (e.g., [Linde & Labov 75; Ehrlich & Johnson-Laird 82]). We discuss here how visual, spatial, and linguistic knowledge structures can be combined with one another to accomplish the task of incremental route descriptions. Therefore, a flow of information is followed from visual perception towards natural language. The advantage of three representation levels is that there is still a clear distinction between perceptual and linguistic processes and representations. This is in particular important in computational models for distinguishing between spatial relations on the conceptual-spatial level and spatial prepositions on the linguistic level. As a domain for investigating the relation between visual perception, natural language, and intermediate processes and representations, we use incremental route descriptions. Route descriptions can be distinguished into *complete* and *incremental route descriptions* (cf. [Maaß 93]). Incremental route descriptions are given step-by-step while moving along the path towards the destination, as from a co-driver. Hence, incremental route descriptions in combination with processes of visual perception are ideal for investigating different representation levels of spatial information. Complete route descriptions are given in advance by using spatial knowledge stored in long-term memory, which generally relates to research about 'cognitive maps' (cf. [Lynch 60; Downs & Stea 73; Allen & Kirasic 85; Hirtle & Jonides 85; McNamara et al. 92; Tversky 92]). Research in this domain is primarily interested in how people represent and retrieve spatial information. The uniform linguistic structure of German route descriptions is the reason why syntactic and semantic structures of route descriptions have been investigated by several linguistic studies (e.g., [Klein 82; Wunderlich & Reinelt 82; Habel 87; Meier et al. 88; Hoeppner et al. 90]). A comparison of complete and incremental route descriptions shows that in the incremental case linguistic structures depend more on descriptions of actions.

But a viewer/speaker[5] has the additional tasks of moving through and anticipating changes in the environment. What has not been generally considered in this context are temporal dependencies. We outline how temporal dependencies are integrated in the proposed computational model to achieve adaptive and appropriate behavior.

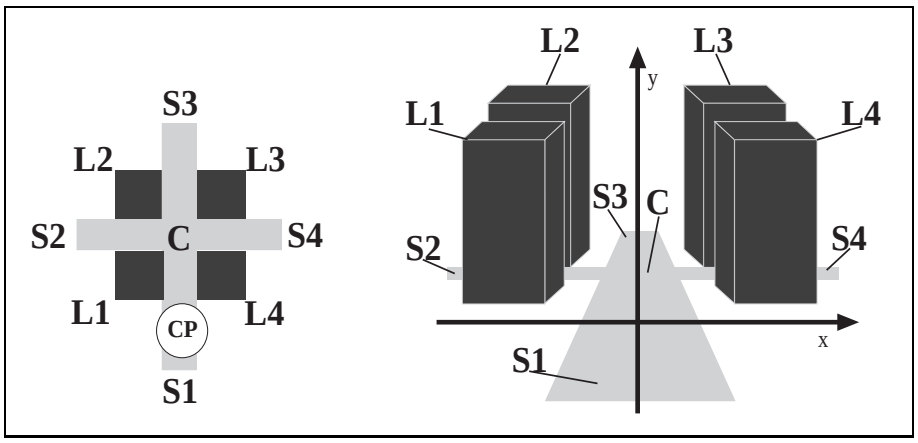

Fig. 1. 2D- and 3D view of a crossing

In the proposed model we distinguish between three different types of objects: a person (unity of viewer and speaker), street items (street segments, decision points), and landmarks (buildings, cars, trees, signs, etc.). Each street item and each landmark are related to MOSES by one spatial relation. MOSES can always describe the position of a visible item in relation to his egocentric frame of reference. The motivation for the distinction between objects and spatial relations is that objects do not appear to "fly" around in space. If we perceive a situation, objects are spatially related to one another. In most models spatial relations between objects are defined on the basis of coordinates in an Euclidean

[5] Before you can give a incremental route description you must visually obtain information from the environment. Therefore, MOSES is a combination of a viewer and a speaker.

system (e.g., [Müller 88; Gopal et al. 89; Hoeppner et al. 90]). In these models it is taken for granted that exact positions of objects are provided. Therefore, they are closely related to Geographical Information Systems (cf. [Frank 87; Goodchild 88]). This does not, however, seem to reflect how object locations are represented by the human mind. Research about 'cognitive maps' indicates that mental representations of space are quite inaccurate (e.g., [Tversky 92]), either because the representations themselves are fuzzy or because inference processes on these representations are not as exact as coordinates. From an efficiency perspective it is unreasonable to assume that we first obtain highly accurate geometric information and then transform this during subsequent steps into fuzzy long-term representations. A complementary approach is to use representations based on qualitative spatial relations. Kuipers and Freksa, for instance, propose mechanisms for interrelating places, streets, and the viewer to one another by qualitative spatial structures (cf. [Kuipers 78; Freksa 91]).

We asked people in computer-simulated and real-world environments to give incremental route descriptions. Similar to the results for complete route descriptions, we found that the structure of incremental route descriptions are quite schematic. The schematic structure is important for the process model here. These findings strongly relate to Neisser's visual perception cycle (cf. [Neisser 76]) and his use of schemata, Johnson-Lairds mental models (cf. [Johnson-Laird 83]), and Herrmann's *HOW schemata* (cf. [Herrmann & Grabowski 94]). In AI there are several approaches for formalizing the idea of schemata, such as Minsky's FRAMES or Schank and Abelson's SCRIPTS (cf. [Minsky 75; Schank & Abelson 77]). Schemata are *compiled knowledge* about generally limited domains. FRAMES and SCRIPTS provide a framework for expectations which represent situations compatible with the structure of the domain. The use of schemata is only appropriate in domains with clear-cut structures.

2 Towards a computational model

Central to our computational model is a situated agent, called MOSES, who moves through simulated 3D environments (for details see [Maaß 93; Maaß 94]). MOSES selects a path from a map. His task is to describe this path and the environment step-by-step to a listener, who is assumed to follow him (see figure 2 for a view on the graphical user interface[6]). This can be metaphorically described as a driver co-driver scenario (for a review of different computational models for navigation refer to [Maaß 94]). At the linguistic level, spatial knowledge is transformed into linguistic knowledge structures. We first describe how information at the spatial level is constructed and modified, followed by a description of the transformation process.

Following Marr, we assume that the visual system generates 3D representations of items obtained from the environment. How we construct 3D representations is, however, beyond the scope of this article (for details see [Herzog et al. 89; Koller et al. 92; Rohr 94]). MOSES has 3D-representations for different types of objects, such as buildings, streets, and cars. The interesting point is which objects and relations are selected from a input stream of visual information. We have determined a computational model for selecting objects by *visual salience* which is based on Treisman's *feature integration theory* (cf. [Maaß 95b]). Visual features, such as color, size, direction of movement, and orientation, are grouped into *feature clusters*. Only those entities which are 'indexed' by features and feature clusters are considered for the identification and categorization of objects. Path-related intentions which determine whether to turn right, left, or to go straight on at the next decision point guide MOSES' focus of *spatial attention area* (see figure 3). Items which lie in the spatial attention area are preferred. If an entity is considered to be salient within a given context it is identified by

[6] The current version of MOSES is implemented in CommonLisp and CLOS with the graphical user interface written in CLIM. The system has been completely developed on Hewlett Packard Series 700 and SPARC workstations

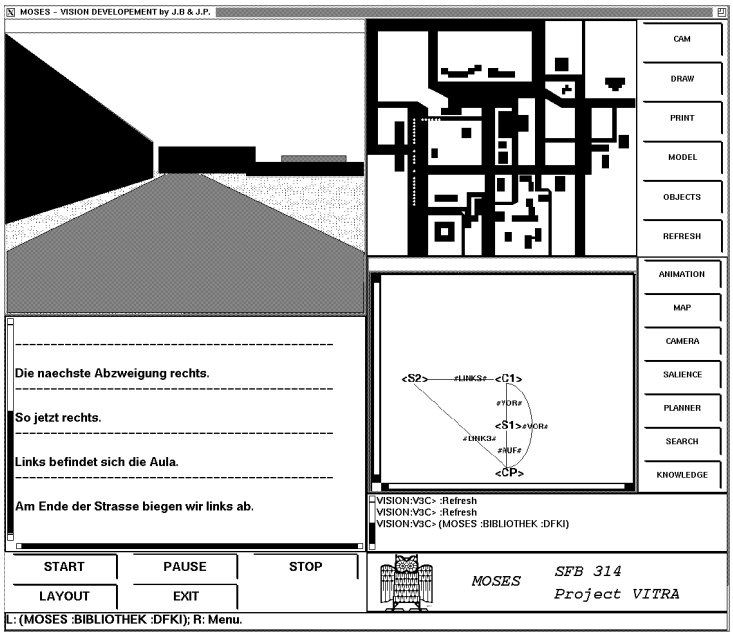

Fig. 2. A View on MOSES

matching it with object schemata (for more details see [Maaß 95b]).

Once the objects have been selected, a set of spatial relations between them is determined. Therefore, MOSES transforms the perspective view of a situation into a two-dimensional representation adopting a top-down view. Objects and corresponding spatial relations are integrated in a coherent structure, called a *configuration description*. Here objects are related to one another and to MOSES by geometric spatial relations. Configurational descriptions, which are networks of spatial relations between objects (MOSES' egocentric frame of reference, landmarks, and street items), are divided into two categories: *minimal* and *extended configuration descriptions*. As will be described later, this distinction is mainly motivated by the consideration of temporal and situative constraints. Minimal configurational descriptions only include MOSES' location, street items, and the spatial relations between them. Hence, a minimal configurational description is

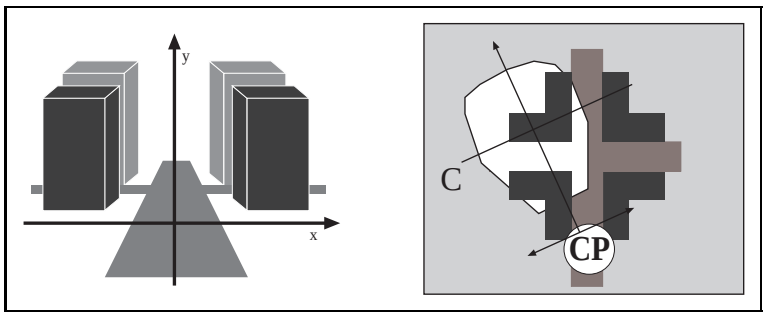

Fig. 3. Focus of spatial attention – top-down and perspective view

the minimal amount of information required about the environment which enables MOSES to follow the path. If there is enough time, MOSES also determines landmark information, i.e., he selects landmarks which can be used for describing the next action. If a landmark is integrated in the configurational description we call it *extended configuration description*. For representations on the conceptual-spatial level, we use a restricted set of binary spatial relations, i.e. #left-of#, #right-of#, #in-front-of# and #behind#[7], but obviously not all possible spatial relations between objects, street items, and the viewer are actually determined. As indicated by figure 4, a complex configuration emerges if MOSES only determines one spatial relation between MOSES (CP), street items and landmarks, between landmarks and nearest street items, and between connected street segments and decision points[8].

We asked people to describe turn actions in computer-animated crossing scenes. We found that in time-restricted situations people tended to limit the length of their descriptions. If they had enough time they also referred to salient

[7] The # indicates that these spatial relations are distinct from spatial prepositions, such as "left of". At the moment it is unclear whether the type of listed set of conceptual spatial relations is appropriate, but it is quite obvious that the four relations are not sufficient to represent all configurations.

[8] A decision point is a location on a street where the viewer has to decide how to continue a path. At a decision point MOSES might turn left or right or go straight on.

landmarks. Two classes of spatial relations can be distinguished. First, all street items are related to MOSES' egocentric frame of reference. Second, street segments and decision points are related to one another. Street items, such as street segments and decision points are of primary interest. Landmarks do not provide important information for following a path, whereas without a proper representation of street information, MOSES is not able to follow a path. There is a difference between directly accessible relations and those which must be inferred. For instance, in figure 1 the relation between S2[9] and L4 is not as easy to describe as the relation between S2 and C. In reference to MOSES' location, S2 is in the left half plane and L4 in the right one. Furthermore, the distance between S2 and L4 is greater than the distance between S2 and C. We say that S2 and L4 are not *visually near* to one another (see [Maaß 95a]). Two objects are visually near if they share the same visual area on the projection plane (see figure 1). For instance, in the perspective view of the crossing (see figure 1), L1, S2, and L2 share a similar area on the projection plane, i.e. these objects are visually near.[10]

For efficiency reasons, MOSES only evaluates a minimal set of spatial relations. It is inefficient to evaluate all spatial relations in every situation, especially in the case of moving objects. Therefore, a procedure incrementally adjusts the configuration description to the environment. When a new configuration is formed, first of all the spatial relations between all items and the viewer need to be determined. Then spatial relations between street items can be established. The resulting structure is called a *minimal configurational description*. When MOSES selects a new landmark it is first related to MOSES' egocentric frame of reference by computing the best applicable spatial relation between MOSES

[9] CP is the current position of MOSES, L1 to L4 are landmarks, S1 to S4 are street segments, and C is the crossing section as indicated in figure 1.

[10] At the moment we do not consider depth information and experience of the viewer. Currently the distinction into a left and a right visual plane is important for the determination of visual nearness. The next step is to evaluate whether visual nearness must also refer to depth information.

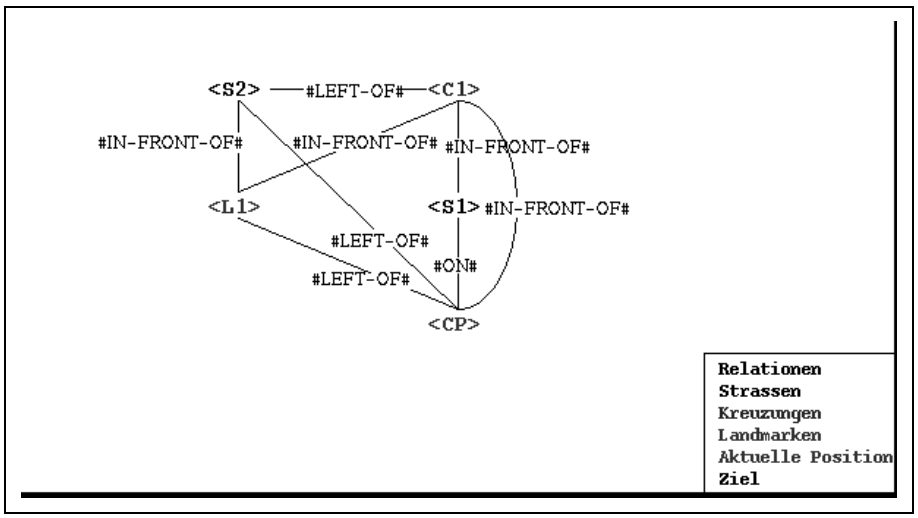

Fig. 4. Example for an extended configurational description

and the landmark. The next step is to establish all spatial relations between the landmark and visually near items. Finally, a landmark or a street item is deleted if it is no longer visible or MOSES turns into a new street segment (cf. [Gopal et al. 89] for an initial approach to modeling the decay of spatial knowledge). If landmark information is selected from the environment, the representation structure is called an *extended configurational description* (see landmark L1 in figure 4). In time-restricted situations, MOSES is forced to depend on a minimal amount of information obtained from the environment. Therefore the information represented by the minimal configuration is the basics for MOSES to be able to orientate himself in complex environments. A spatial relation between MOSES's current position and the current street segment S1 is defined by the spatial relation #ON#(CP, S1). In the same way, the crossing C and street segment S2 are related to MOSES (see figure 4) by: (#IN-FRONT-OF#(C, CP) $\wedge$ (#LEFT-OF#(S2, CP)). Besides these relations, MOSES determines a minimal set of relations between street items. In order to avoid a combinatorial explo-

sion not all possible spatial relations between street items are evaluated, only those between physically connected street items, e.g., (#LEFT-OF#(S2, C) $\wedge$ (#IN-FRONT-OF#(S1, C)). In MOSES we have a set of configuration schemata for a sample set of decision point situations with one additional landmark and procedures for combining schemata.

In summary, if an object is salient in a given situation it will be identified by a visual selection process. This triggers a process which integrates this landmark into a configurational description by determining geometric spatial relations between MOSES' current position, street items, and selected landmarks. As we will show next, configurational descriptions are important for the determination of approriate incremental route descriptions.

2.1 Selection of description schemata

We have already mentioned that spatial relations used in configuration descriptions are an initial approach, and mainly coined by verbal descriptions collected by our experiments. Now we describe how configurational descriptions are matched with linguistic structures. This mainly depends on findings that the linguistic structure of German route descriptions is schematic (cf. [Klein 82; Habel 87; Wunderlich & Reinelt 82; Meier et al. 88; Müller 88; Hoeppner et al. 90]). In familiar urban environments, we depend on experience and schemata about how particular objects are expected to be distributed in space. For instance, if we reach a crossing we expect to see buildings on the left and on right hand side and a street going in-between (New York is a master example for that). A ship in the middle of the crossing would cause us to hesitate because it does not fit into our general expectations about traffic situations. By experiments in computer-simulated and real world environments we collected a corpus of incremental route descriptions. In the first experiment, we asked test persons to describe turn left, turn right or go-straight actions in a computer simulated crossing scenario (the scenes presented in these experiments have been

similar to the one presented in figure 1). In the first setting a simulated car was driven through an environment with medium speed. Most test persons only described the turning action itself. In settings with lower speed the test persons also included salient landmarks in their descriptions. In settings where they were asked to include a particular landmark they had difficulties in giving a correct description when the landmark was on the opposite side of the y-axis at the next street segment (see figure 1). In this setting most test persons described that the action ("An der nächsten Kreuzung biegst du links ab." ["At the next crossing turn left."]) followed by an extension of the description ("... dort, am ersten Gebäude auf der rechten Seite." ["... there, by the first building on the right-hand side."]). Some persons were not even able to integrate the indicated landmark. One possible conclusion is that in the second setting the landmark on the opposite side does not fit into the preferred schema of describing a situation. By examining the corpus of descriptions, we found that most descriptions can be categorized by a small set of syntactic schemata (for details see [Maaß 95a]). In particular, the categorization into 'WHAT', 'WHEN', 'WHERE', and 'WHERE TO'-phrases is helpful in understanding the structure of route descriptions. A 'WHAT'-phrase describes an action and is usually a verb phrase, e.g., "... mußt du abbiegen ..." (... you must turn ...). Temporal descriptions are introduced by 'WHEN'-phrases, e.g., "... jetzt ..." (... now ...). 'WHERE'-phrases describe the location of a landmark or a location where an action must be performed, e.g., "Da vorne ..." ["There in front ..."] or "Zwischen den beiden Häusern ..." ["Between those two buildings ..."]. An extension of a 'WHERE' phrase is a 'WHERE TO' phrase. The direction of an action is indicated by referring to locative information, e.g., "... nach links ..." ["... left ..."]. 'WHERE TO' phrases are commonly connected to 'WHAT' phrases, e.g., "... nach rechts abbiegen ..." ["... turn right ..."]. For instance, a typical description is: "Bitte gleich rechts abbiegen... hinter dem braunen Gebäude... Jetzt bitte" ["Please turn right ... after the brown building ... Now please."]. The structure of

this corpus of German descriptions[11] can be described as a sequence of WHAT-WHERE-WHEN-WHERE TO phrases. We found that the test persons used in 70 percent of cases, one of the following phrase structures: WHERE+WHERE TO+WHAT, WHERE+WHERE TO, or simply WHERE. Based on these sequences we extracted a set of linguistic schemata, called *description schemata*. On the one hand, they reflect the linguistic structure of route descriptions and on the other hand, they correspond to spatial information represented by configuration descriptions. A configurational description provides explicit information about the spatial structure of a situation. Route descriptions mainly depend on the spatial structure represented by configurational descriptions. MOSES considers the given configurational description, intentions, the temporal structure of the situation, his linguistic abilities and knowledge about the listener to select an appropriate description schema (see figure 5). The temporal structure of a situation is constrained by the speed of MOSES and the distance to the next decision point. MOSES makes assumptions about how long it will probably take to reach the next decision point. According to this time interval, only those schemata which can be used to generate a description in time are selected. The next filter selects from these schemata those which correspond to the intended action at the next decision point. For this, only simple path-related intentions are considered, i.e. intentions to turn right, to turn left, or to go straight on. During the next selection step, those schemata are extracted which assume a similar spatial structure to that given by the configurational description. If there are objects selected by the object selection process, then those schemata are prefered which include a reference to salient objects at appropriate places. Most of all, MOSES descriptions depend on his type of movement. When he moves for instance with average car speed, intervals between decision points are sometimes quite short.

[11] It is interesting to note that the phrases in the corresponding English descriptions are very similar in their structure. However, our corpus exceptionally consists of German descriptions. Hence we cannot draw any conclusions for other languages, although it seems that there are strong correlations.

In these situations, he only refers to route knowledge. If he moves at walking speed, he has more time and can refer to objects. For instance, if a salient object is on the right and his intention is to turn right he gives the description in two parts: "Please, turn right after the red building on the right." .... "Now, turn right please." First, he gives a complete description of the intended action by referring to objects. Then, just before the action needs to be accomplished, he gives an additional hint. During the last two selection steps those schemata are selected which correspond to the properties of the speaker and the listener (for more details see [Maaß 95a]). This selection process extracts and instantiates one or more description schemata. If there are more than one schemata, MOSES uses the first one. It is clear that a more sophisticated conflict resolution procedure would be helpful, but in our domain we found that this simple strategy serves quite well.

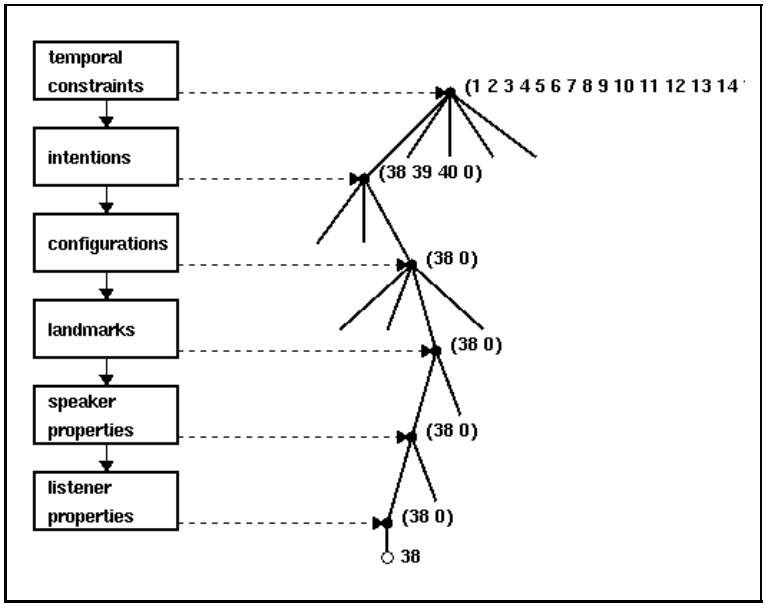

Fig. 5. Selection of a description schema

A description schema represents the semantics of a particular incremental route description. The structure of a schema is based on Jackendoff's *conceptual semantics* (cf. [Jackendoff 83]) and because these are based on simple utterances we carefully extended his formalism (for an example see figure 6). MOSES has a repertoire of almost 60 description schemata. Basic constituents of a description schema are *things* (persons), *locations* (places), and *paths*. They are used in higher-order structures, such as *events* and *states*. The general structure of an event consists of a reference to the listener's reference frame followed by a description of a path and a place. Hence we can represent utterances such as: "Please, turn right after the building on the right." Figure 6 shows the conceptual structure of this description.

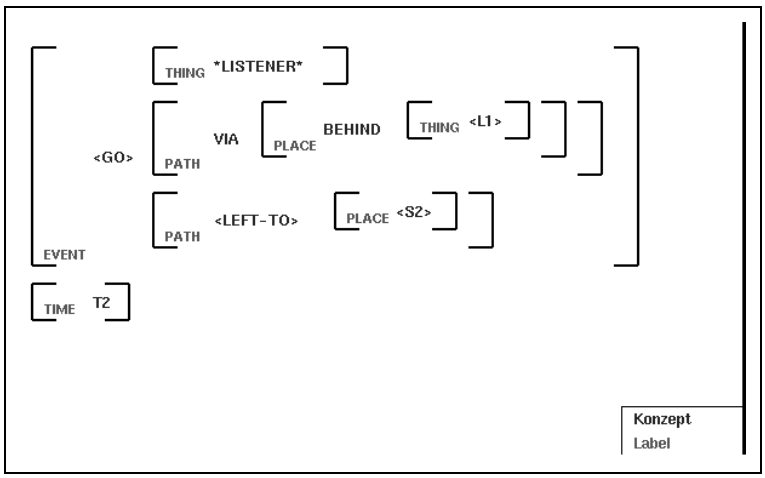

Fig. 6. Example for an *event* description schema

The marker <GO> indicates that schema represents a description of an action. The description is given by adopting the listener's perspective, i.e., the listener's egocentric frame of reference. Then two substructures of type *path* follow. L1 is the pointer to the structure of this landmark. If this location behind

L1 is reached, the listener must turn left into a place referred to by S2. Finally, the temporal marker indicates the temporal interval (t2) when the description must be given. These time intervals (t1 to t5) are extracted from experiments in real environments, where we found that five intervals can be distinguished. Most descriptions are given during the last 10 seconds before a action need to be performed. Two time intervals correspond to this time interval (t4 and t5). The environment are presently restricted to crossings with rectangular street configurations. In the future, however, this will be extended so that more complex decision point configurations can also be described. Then, a description schema is transformed into surface structures which serves as input for the natural language process. (For a description of input structure for the natural language generator see [Maaß 95b]). The visual object selection process provides information about landmarks, such as color, height, and width. This kind of information is used for referring to physical attributes of landmarks (cf. [Maaß et al. 95]). Finally, MOSES generates the following description: "Biegen Sie hinter dem günen Haus links ab." ["Turn left after the green building on the left-hand side."].

2.2 Adaptation to the environment

Most AI systems are only built to do *something*. A recently emerging constraint is that they are also required to do something *at anytime*. For instance, robots should not stop on railway tracks to reason about what to do next. An important constraint of anytime algorithms is that the quality of behavior increases with the quantity of the limited resource (cf. [Russel & Wefald 91]). Relating this to the domain of route descriptions means that a description should asymptotically increase in quality with the available amount of time. Most models dealing with spatial knowledge have the basic assumption that the processing time is small compared with the numbers of events in the environment. On the one hand, temporal constraints are established by physical events in the environment. On the other hand, and more important for models of cognitive processes, temporal

constraints are subjectively measured by the agent. When MOSES approaches a decision point he makes assumptions about how long it will probably take to reach this point. This is a basic temporal constraint to which other processes refer. By measuring time intervals, we found that people who where asked to incrementally describe a route in a real environment showed a common pattern. In situations where the next decision point was far away, test persons tended to give the description of the next action about 10 seconds before arriving at the decision point. In some cases the test person explicitly mentioned that he/she had waited to give the description at the 'right' point of time. In situations where the next decision point was quite near, he/she reduced the complexity of the description, i.e. by only referring to street items. This motivates the distinction between minimal and extended configuration descriptions. First the minimal configuration is generated and used as input for the language generation system. If landmark information and additional path information is obtained, MOSES extends the minimal configuration. This allows MOSES to describe a situation after a short initialization phase. Because MOSES moves through a simulated environment, he adjusts his descriptions to his own movements and to changes in the environment. In situations with little time he only selects and describes a restricted set of visual items. For instance, if MOSES turns left at a crossing and the time interval to the next decision point is only about 10 seconds, then he does not have enough time to analyze the whole scene in detail. Therefore, the description is adapted to this temporal limitation. The main task is to give the appropriate descriptions at the 'right' point of time so that the listener knows where to perform which kind of actions.

3 Summary and conclusion

Incremental route descriptions are ideal for investigating the representation levels of spatial knowledge. We have outlined a three-level approach for representing spatial information consisting of a visual level, a conceptual-spatial level,

and a linguistic level. We focused on the interrelation of representations on the conceptual-spatial level and the linguistic level. Incremental route descriptions provide a well-structured domain for the investigation of the distinction between these three levels. Fundamental to the model is the dissection into mode-specific and mode-independent representations of space. Spatial information obtained by visual perception is represented by 3D models, but there is evidence to assume a mode-independent representation structure on a conceptual-spatial level between visual perception and natural language. Spatial information about the environment, which is stored in configurational descriptions, is used as input for a description schema selection procedure. Central to MOSES is the inherent schematic structure of incremental route descriptions. What has generally not been considered up to now in the context of route descriptions are the influence of temporal constraints. Therefore, we indicated the importance of temporal constraints, as well as their integration into MOSES to achieve 'anytime' behavior. We are currently working on a model for temporal constraints and how it affects the 'anytime' behavior of MOSES. Furthermore the connection between presentation schemata and language generation is examined in detail. We are also investigating the hierarchical organization of configuration descriptions and description schemata. By our corpus of route descriptions and further experiments, we hope to gain more insights into these structures.

Since visual perception and natural language are two complex research areas on their own, we are far from having anything like a complete theory which will integrate both fields. Nevertheless further efforts focusing on the integration of both areas are required for a better understanding of cognitive processes and representations and also for their use in computational systems.

4 Acknowledgements

I would like to thank Jörg Baus and Joachim Paul for taking care of the implementation issues and for fruitful discussions. In addition, for comments from

reviewers which helped to improve the quality of this article in many ways. This work is supported by a grant of the Graduiertenkolleg *Kognitionswissenschaft* at the University of the Saarlandes, Saarbrücken.

References

[Allen & Kirasic 85] G. L. **Allen** and K. C. **Kirasic**. *Effects of the cognitive organization of route knowledge on judgments of macrospatial distance.* Memory and Cognition, 13(3):218–227, 1985.

[André et al. 89] E. **André**, G. **Herzog**, and T. **Rist**. *Natural Language Access to Visual Data: Dealing with Space and Movement.* In: F. Nef and M. Borillo (eds.), Logical Semantics of Time, Space and Movement in Natural Language. Proc. of 1^{st} Workshop. Hermès, 1989.

[Baddeley & Hitch 74] A. D. **Baddeley** and G. J. **Hitch**. *Working Memory.* In: G. Bower (ed.), Recent advances in learning and motivation. New York: Academic Press, 1974. Vol. VIII.

[Baddeley 86] A. D. **Baddeley**. *Working Memory.* Oxford: Oxford University Press, 1986.

[Binford 71] T. O. **Binford**. *Visual Perception by Computer.* In: Proc. IEEE Conf. on Systems and Control, 1971.

[Bryant 92] D. J. **Bryant**. *A Spatial Representation System in Humans.* Journal of Memory and Language, 31:74–98, 1992.

[Couclelis 94] H. **Couclelis**. *Verbal directions for way-finding: space, cognition, and language.* technical report, Department of Geography, UC Santa Barbara, 1994.

[Downs & Stea 73] R. M. **Downs** and D. **Stea**. *Cognitive Maps and Spatial Behaviour: Process and Products.* In: R. M. Downs and D. Stea (eds.), Image and Environment. Cognitive Mapping and Spatial Behaviour, pp. 8–26. Chicago: Aldine, 1973.

[Ehrlich & Johnson-Laird 82] K. **Ehrlich** and J. N. **Johnson-Laird**. *Spatial descriptions and referential continuity.* Journal of Verbal Learning and Verbal Behavior, 21:296–306, 1982.

[Frank 87] A. **Frank**. *Towards a Spatial Theory.* In: Proc. of the International Symposium on Geographic Information Systems: The Research Agenda, pp. 2:215–227, Crystal City, Virginia, 1987.

[Freksa 91] Ch. **Freksa**. *Conceptual neighborhood and its role in temporal and spatial reasoning.* In: M. Singh and L. Trave-Massuyes (eds.), Decision support systems and qualitative reasoning, pp. 181–187. Amsterdam: North-Holland, 1991.

[Goodchild 88] M. F. **Goodchild**. *Towards an Enumeration and Classification of GIS Functions.* In: Proc. of the International Geographic Information Systems Conference: The Research Agenda, pp. II:67–77, Washington, 1988. NASA.

[Gopal et al. 89] S. **Gopal**, R. **Klatzky**, and T. **Smith**. *NAVIGATOR: A Psychologically Based Model of Environmental Learning Through Navigation.* Journal of Environmental Psychology, 9:309–331, 1989.

[Habel 87] Ch. **Habel**. *Prozedurale Aspekte der Wegplanung und Wegbeschreibung.* LILOG-Report 17, IBM, Stuttgart, 1987.

[Herrmann & Grabowski 94] T. **Herrmann** and J. **Grabowski**. *Sprechen: Psychologie der Sprachproduktion.* Spektrum, Akademischer Verlag, 1994.

[Herskovits 86] A. **Herskovits**. *Language and Spatial Cognition. An Interdisciplinary Study of the Prepositions in English.* Cambridge, London: Cambridge University Press, 1986.

[Herzog et al. 89] G. **Herzog**, C.-K. **Sung**, E. **André**, W. **Enkelmann**, H.-H. **Nagel**, T. **Rist**, W. **Wahlster**, and G. **Zimmermann**. *Incremental Natural Language Description of Dynamic Imagery.* In: Ch. Freksa and W. Brauer (eds.), Wissensbasierte Systeme. 3. Internationaler GI-Kongreß, pp. 153–162. Berlin, Heidelberg: Springer, 1989.

[Hirtle & Jonides 85] S. **Hirtle** and J. **Jonides**. *Evidence of hierarchies in cognitive maps.* Memory and Cognition, 13(3):208–217, 1985.

[Hoeppner et al. 90] W. **Hoeppner**, M. **Carstensen**, and U. **Rhein**. *Wegauskünfte: Die Interdependenz von Such- und Beschreibungsprozessen.* In: C. Freksa and C. Habel (eds.), Informatik Fachberichte 245, pp. 221–234. Springer, 1990.

[Jackendoff 83] R. **Jackendoff**. *Semantics and Cognition.* Cambridge, MA: MIT Press, 1983.

[Johnson-Laird 83] P. N. **Johnson-Laird**. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness.* Cambridge University Press, 1983.

[Klein 82] W. **Klein**. *Local Deixis in Route Directions.* In: R. J. Jarvella and W. Klein (eds.), Speech, Place, and Action, pp. 161–182. Chichester: Wiley, 1982.

[Koller et al. 92] D. **Koller**, K. **Daniilidis**, K. **Thórhallson**, and H. H. **Nagel**. *Model-based Object Tracking in Traffic Scenes.* In: G. Sandini (ed.), The Second European Conference on Computer Vision, pp. 437–452, Berlin, Heidelberg, 1992. Springer.

[Kuipers 78] B. **Kuipers**. *Modeling Spatial Knowledge.* Cognitive Science, 2:129–153, 1978.

[Lakoff 87] G. **Lakoff**. *Women, Fire, and Dangerous Things. What Categories Reveal about the Mind.* Chicago: Chicago University Press, 1987.

[Linde & Labov 75] C. **Linde** and W. **Labov**. *Spatial Network as a Site for the Study of Language and Thought.* Language, 51:924–939, 1975.

[Lynch 60] K. **Lynch**. *The Image of the City.* MIT Press, 1960.

[Maaß et al. 95] W. **Maaß**, Jörg **Baus**, and Joachim **Paul**. *Visual Grounding of Route Descriptions in Dynamic Environments.* In: AAAI Fall Symposium on "Computational Models for Integrating Language and Vision", MIT, Cambridge, MA, 1995. AAAI. in print.

[Maaß 93] W. **Maaß**. *A Cognitive Model for the Process of Multimodal, Incremental Route Description.* In: Proc. of the European Conference on Spatial Information Theory. Springer, 1993.

[Maaß 94] W. **Maaß**. *From Visual Perception to Multimodal Communication: Incremental Route Descriptions.* Artificial Intelligence Review Journal, 8(5/6), December 1994. Special Volume on Integration of Natural Language and Vision Processing.

[Maaß 95a] W. **Maaß**. *Generierung multimodaler inkrementeller Wegbeschreibungen in dynamischen 3D-Umgebungen.* PhD thesis, Universität des Saarlandes, 1995. in preparation.

[Maaß 95b] W. **Maaß**. *Selection of objects by evaluation of visual features.* in preparation, 1995.

[Marr & Nishihara 78] D. **Marr** and H. K. **Nishihara**. *Representation and Recognition of the Spatial Organization of three-dimensional shapes.* In: Proc. Royal Society of London B, pp. 269–294, 1978.

[Marr 82] D. **Marr**. *Vision: a computational investigation into the human representation and processing of visual information.* San Francisco: Freemann, 1982.

[McKevitt 94a] P. **McKevitt** (ed.). *Integration of Natural Language and Vision Processing.* AAAI-94 Workshop. Seattle, WA, 1994.

[McKevitt 94b] P. **McKevitt** (ed.). *Special Volume on the Integration of Natural Language and Vision Processing*, volume 8: Artificial Intelligence Review Journal. Dordrecht: Kluwer, 1994.

[McNamara et al. 92] T. **McNamara**, J. **Halpin**, and J. **Hardy**. *The representation and integration in memory of spatial and nonspatial information.* Memory and Cognition, 20(5):519–532, 1992.

[Meier et al. 88] J. **Meier**, D. **Metzing**, T. **Polzin**, P. **Ruhrberg**, H. **Rutz**, und M. **Vollmer**. *Generierung von Wegbeschreibungen.* KoLiBri Arbeitsbericht 9, Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, 1988.

[Minsky 75] M. **Minsky**. *A Framework for Representing Knowledge.* In: P. H. Winston (ed.), The Psychology of Computer Vision. New York: McGraw-Hill, 1975.

[Müller 88] S. **Müller**. *CITYGUIDE: Ein System zur Wegplanung und Wegbeschreibung.* Diplomarbeit, Fachbereich Informatik der Universität des Saarlandes, 1988.

[Neisser 76] U. **Neisser**. *Cognition and Reality.* San Francisco: Freeman, 1976.

[Rohr 94] K. **Rohr**. *Towards Model-based Recognition of Human Movements in Image Sequences.* Computer Vision, Graphics, and Image Processing (CVGIP): Image Understanding, 59(1):94–115, 1994.

[Russel & Wefald 91] S. **Russel** and E. **Wefald**. *Do the Right Thing: Studies in Limited Rationality.* Cambridge, MA: MIT Press, 1991.

[Schank & Abelson 77] R. C. **Schank** and R. P. **Abelson**. *Scripts, Plans, Goals and Understanding.* Hillsdale, NJ: Erlbaum, 1977.

[Schirra et al. 87] J. R. J. **Schirra**, G. **Bosch**, C.-K. **Sung**, and G. **Zimmermann**. *From Image Sequences to Natural Language: A First Step Towards Automatic Perception and Description of Motions.* Applied Artificial Intelligence, 1:287–305, 1987.

[Talmy 83] L. **Talmy**. *How Language Structures Space.* In: H. Pick and L. Acredolo (eds.), Spatial Orientation: Theory, Research and Application, pp. 225–282. New York, London: Plenum, 1983.

[Tversky 92] B. **Tversky**. *Distortions in cognitive maps.* Geoforum, 23:131–138, 1992.

[Wunderlich & Reinelt 82] D. **Wunderlich** and R. **Reinelt**. *How to Get There From Here.* In: R. J. Jarvella and W. Klein (eds.), Speech, Place, and Action, pp. 183–201. Chichester: Wiley, 1982.

This article was processed using the LaTeX macro package with LLNCS style