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. 108

From Visual Perception to Multimodal Communication: Incremental Route Descriptions

Wolfgang Maaß

Juli 1994

ISSN 0944-7814

108

From Visual Perception to Multimodal Communication: Incremental Route Descriptions*

Wolfgang Maaß

Cognitive science program
Universität des Saarlandes
D-66041 Saarbrücken
maass@cs.uni-sb.de

Abstract

In the last few years in cognitive science there has been a growing interest in the connection between visual perception and natural language. The question of interest is: How can we discuss what we see. With this question in mind in this article we will look at the area of *incremental route descriptions*. Here, a speaker step-by-step presents the relevant route information in a 3D-environment. The speaker must adjust his/her descriptions to the curenntly visible objects. Two major questions arise in this context: 1. How is visually obtained information used in natural language generation? and 2. How are these modalitites coordinated? We will present a computational framework for the interaction of visual perception and natural language descriptions which integrates several processes and representations. Especially discussed is the interaction between the spatial representation and the presentation representation used for natural language descriptions. We have implemented a prototypical version of the proposed model, called `MOSES`.

*This work is funded by the cognitive science program *Graduiertenkolleg Kognitionswissenschaft* of the German Research Community (DFG)

1 Introduction

In the last few years in cognitive science there has been a growing interest in the connection between visual perception and natural language. The reducionist approach to seperate each of these topics tends to miss the fact that human beings integrate both abilities in order to act and communicate in their environment.

In the project VITRA (Visual Translator), we are investigating the connection between visual perception and multimodal communication in dynamic environments from a computational point of view (cf. Herzog and Wazinski, this volume). In a subproject, we are concerned with the generation of multimodal incremental route descriptions that combine natural language and spontaneous gestures[1] (cf. [Maaß 93]). The central question for this kind of communication is how to lead a person to his/her destination by describing the route during the trip itself. Persons we have asked to give incremental route descriptions have always used different modalities, e.g. speech and spontaneous gestures, to describe the route. Thus, two of the interesting subquestions are: 1. How is visually obtained information used in natural language production? and 2. How are these modalitites coordinated? Our model, called `MOSES`, is led by psychological results (cf. [Allen & Kautz 85]) that humans mentally partition the continuous and complex spatial environment into segments of information. These segments represent a partial view of the actual environment. We propose two selection steps: the visual selection and the presentation selection. These selections reduce the complexity of information and the complexity of computation and thus make it possible to describe the route in an efficient way.

In section 2, we mention some of the major problems which seem to be central for incremental route descriptions followed by a discussion of related research. For the process of incremental route descriptions we have determined a computational model which is presented in section 3. Two different input processes determine the behaviour of the whole process: the visual perception (section 3.1) and the determination of path information obtained by maps (section 3.2). The interaction between the spatial representation and the linguistic representation underlying the presentation processes is presented in section 3.3. In section 4.1, we demonstrate how to plan the presentation structures in order to communicate the obtained information. The modespecific generators for natural language and gestures are

[1]Spontaneous gestures are a beginning research topic (cf. [Sitaro 93]). Sitaro considers spontaneous gestures as a supporting mode for speech. Also it seems that gestures can also carry information on their own. For instance, when we say "Go there!" and we give a spontaneous gesture showing a trajectory rightwards.

briefly mentioned in section 4.2 and in section 5 we give a conclusion and an outlook on some open questions.

2 Motivation

Route descriptions are common communicative actions in everyday life which can be divided into two classes: *complete* (or *pre-trip*) route descriptions and *incremental route descriptions*. In order to give a description of the whole route we use complete route descriptions. Here, a well-known problem for the route finders is that they have to remember a lot of details at one time. On the way to their destination, they normally cannot ask the same person for more details. In *incremental route descriptions*, e.g., given by a co-driver, they receive relevant route information as it is needed. This reduces the cognitive load. In incremental route descriptions, temporal constraints on both the generation and following of route descriptions are central. The construction of a presentation involves, at least the following phases: determination of new information, determination of a presentation structure, transmission of the information, taking into consideration the length of time the hearer will presumably require to understand and verify the information. Furthermore, the information has to be presented in accordance with the strengths of each presentation mode, taking into account the information to be presented and the current environment. In this work, we are addressing the first two phases. In the scenario considered here, the speaker `SP` and hearer `H` travel by car in an urban environment. The task for `SP` is to give adequate information to `H`. With the term *adequate* we mean that `SP` must decide what is relevant to the hearer.

In this project we are not concerned with the use of long-term mental representations of spatial information, often called *cognitive maps*, but rather with the phenomena which arise when the speaker uses a map together with visible information of the current environment in order to determine and to describe the route in an unknown territory. For everyday route descriptions, various kinds of maps are used. The use of maps for description generation, be they street maps, region maps, world maps or anatomic maps, is quite standard. If you want to get from point `A` to point `B` then you first try to orient yourself within the map and in the environment. In the next phase, you look for an appropriate path from `A` to `B`, keeping track of the intended route. Both phases, *orientation* and *wayfinding*, are fundamentally guided by visual perception. Successful navigation requires a consideration of the underlying spatial structure of the environment. On a map of Australia, for instance, types of landmarks and types of paths differ between big cities and the australian outback. In a city, houses and

road signs play an important role whereas in the outback sacred sites and the four cardinal points are the main guidelines because of the lack of other landmarks[2]. In this work, we mainly deal with urban environments.

When we talk about the connection between visual perception and natural language, we have to take into account that both sides usually have different requirements for the representation of spatial information. In the field of visual perception, there have been some approaches to represent shape information about objects. For instance, Marr's 3D model represents an object's shape recursively compounded of cones (cf. [Marr 82]) or Biederman's generalized cone representation (cf. [Biederman 90]). On the linguistic side, representations use geometrical and conceptual information about objects and topological information about path segments as an underlying structure for generating descriptions of spatial arrangements. There are some approaches which try to connect visual perception and natural language production by defining an interface between the spatial representation and the linguistic representation (cf. [Landau & Jackendoff 93, Glasgow 93]).

Psychological experiments have been conducted concerning the human ability to process and represent information about the spatial environment (e.g., cf. [Tolman 48, Piaget et al. 60, Siegel & White 75, Yeap 88]). In the field of Environmental Psychology (cf. [Stokols & Altman 87]), there is a well-known distinction between *route knowledge* and *landmark knowledge* (cf. [Piaget et al. 60, Siegel & White 75]). Route knowledge and its connections to other human abilities have been investigated through several approaches (cf. [Blades 91, Garling 89, Hayes-Roth & Hayes-Roth 79]). In this context, complete route descriptions have been thoroughly examined for psychological (cf. [Streeter et al. 85, Thorndyke & Goldin 83]) and linguistic (cf. [Klein 82, Wunderlich & Reinelt 82, Meier et al. 88, Habel 87]) reasons, but little is known about incremental route descriptions. Klein and Wunderlich have divided the communication of complete route desriptions into four phases: introduction, main part, verification and final phase. The sentence structures in complete route descriptions have been found to be very restricted and sometimes schematic. This is supported by route descriptions in our corpus where we found 31 different descriptions of spatial relations and 13 different verbs of movement. Furthermore, it came out that in general the syntactic structure of each sentence is strongly restricted.

Within the research field of cognitive maps, a distinction is made between *route knowledge* that has more procedural aspects and *survey* or *map knowledge* that has more abstract aspects (cf. [Thorndyke & Hayes-Roth 82, Gluck 91]). In computational models route knowl-

[2]David Lewis has made some interesting investigations concerning the wayfinding ability of aboriginies in the australian outback (cf. [Lewis 76].

edge is more intensively investigated than survey knowledge (cf. [Kuipers 78, McCalla & Schneider 79, Leiser & Zilbershatz 89, Gopal et al. 89]).

3 The proposed computational model

From a phenomenological point of view, we can determine two major phases during the generation of route descriptions. First, the speaker SP looks for an appropriate path from the starting point to the destination which is called *wayfinding*. For this, SP can use different media, most common is the use of maps or mental representations. After the determination of the next path segment, SP describes the route. In general, this description is dominated by speech and spontaneous gestures but also sketches. While SP moves through the dynamic environment and gives descriptions step-by-step SP mentions landmarks, regions and spatial relations obtained by visual perception. This stands in contrast to the human ability to describe a route by only using a cognitive map of a specific area. In this paper, we deal with the phenomenon that the speaker has no experience with the given environment and must use a map to find his/her way.

On the process level, we assume an interaction between the wayfinding and the presentation. Otherwise, it would be hard to explain how we can describe the route together with information obtained by a map or with visible information of the current environment. While describing a non-trivial route, it is quite common to switch between wayfinding and the presentation. Therefore, we assume a process which controls the interaction between both phases.

We have developed a model of the process of *multimodal, incremental route description*. Its general architecture consists of a visual perception process, a wayfinding process, a planning process, and two presentation processes (see figure 1). The wayfinding process models the human ability to look for a route while using a map together with information received by visual perception. Interactions between visual perception, wayfinding and presentation processes are supervised and coordinated by the planning process. In general, all subprocesses depend on temporal aspects concerning the coordination of communication and the current environment, on presumed properties of the questioner, and environmental parameters such as weather, time, and light.

Associated with each process are different representation structures. Visual perception receives information from the current environment. We follow Neisser who proposes that

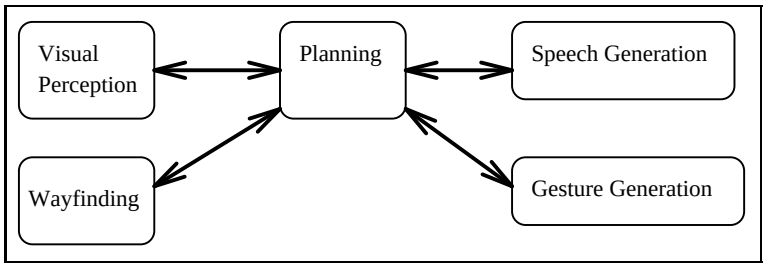

Figure 1: Process Model

the visual perception is led by schemata which enable us to select information even in complex and dynamic environments (cf. [Neisser 76]). We use the term *visual percept* to denote an object-oriented representation structure. It is a dualistic representation consisting of an *object representation* and a *generalized representation* (see figure 3). The object representation integrates a *geometric* and a *conceptual* representation of an object. Landau and Jackendoff investigated that in order to determine spatial relations between objects people use abstract spatial representations of objects (generalized representation). But on the other hand, they found that in order to describe an object verbally we need a geometrical (cf. [Landau & Jackendoff 93]) and a conceptual representation (cf. [Jackendorf 87]).

The wayfinding process extracts path information from a map. In the case of wayfinding, we also use a dualistic representation. On one side, the speaker `SP` has some kind of survey information for the whole path necessary to keep track of a global orientation and to provide abstract information about distance to the destination. On the other side, `SP` matches this general information into an exact topographical partition of the path which leads his/her movement and the description for the current time interval.

Information obtained by visual perception and maps is integrated by the planning process into a central spatial representation structure, called *segment* (see figure 2). Because of the incrementality of the whole process this structure is extended step-by-step. Using this representation, the planning process extracts object information and interrelates relevant objects with one another in order to get a linguistic-oriented presentation structure. A segment provides information about objects, path segments, and spatial relations which are selected for the *presentation representation*. But because of the incrementality, the presentation representation can influence the spatial representation. In order to describe an object uniquely the presentation planner can ask for more information than is already available by the spatial represention. In this case the visual perception must focus on this

object in order to determine more information. The same holds for path information. When the next path segments are needed the wayfinding process searches for new path information.

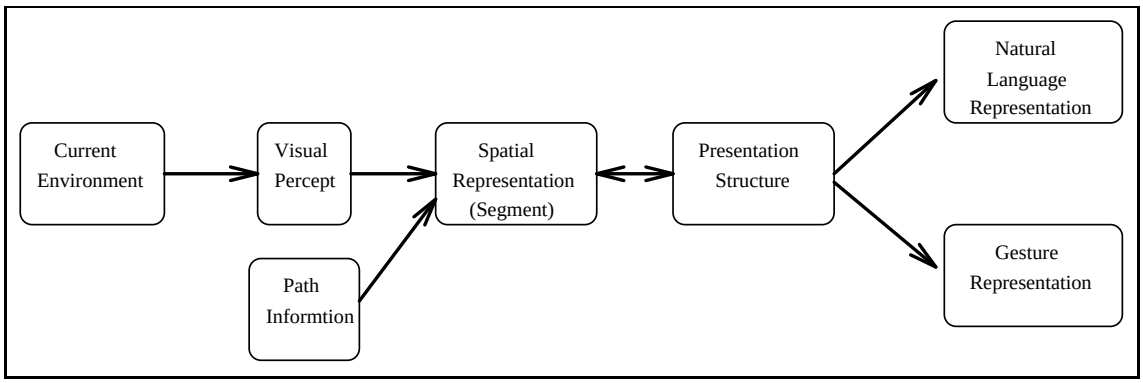

Figure 2: Structure Model

Because our interviews have shown that the linguistic structure of route descriptions is very limited, we use a schematic approach to present relevant information about the environment. These schemata disassociate presentation information into parts for natural language generation and for spontaneous gesture generation (see section4.1).

3.1 What is the role of visual perception in our model?

For our approach, we do not deal with low-level aspects of visual perception (cf. [Marr 82]) but we start on a higher level where specific areas in the current visual field are focussed on. This is similar to the focussing technique used in linguistically oriented approaches (e.g., [Grosz 81]). The current focus defines an *inner* visual field, in which we can identify, categorize and select distinct objects. The current visual focus determines the main landmarks which are used in the presentation: Visual perception leads communication. If the visually accessed information is not sufficient, visual perception will be led by the demands of the presentation. Thus, we want to know: How can high-level visual data be processed efficiently for use in multimodal communication? Our approach is based on psychological investigations by Ittelson (cf. [Ittelson 60]), Neisser (cf. [Neisser 76]), and Schneider/Shiffrin (cf. [Schneider & Shiffrin 77]). Ittelson has performed a number of experiments concerning visual space cues, whereas Neisser has developed a schema-based approach to visual perception. Schneider and Shiffrin found that visual search can be divided into *automatic* and *controlled processes*

Considering these results, the visual perception process generates two kinds of spatial represenation of landmarks: *an object representation* and a *generalized representation.*

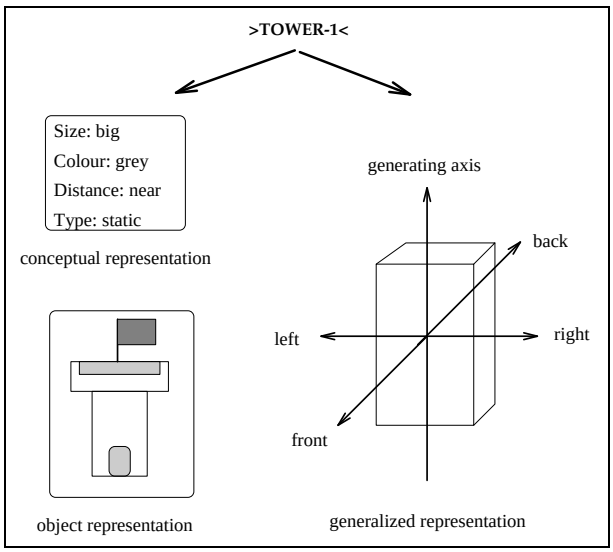

Figure 3: The object representation and generalized representation of landmark `Tower-1`. The generating axis is the vertical axis whereas the orienting axes are horizontal.

In our model, we use cuboids and rectangles to generate geometric representations. Associated with geometric properties the model of a landmark is the conceptual model which includes size, colour, distance to the speaker's position, and functional properties (cf. [Gapp 94, Maaß et al. 94]). The generalized representation is a cuboid which surrounds the landmark associated with two kinds of axis: a *generating axis* (eg., top/down, see figure 3) and *orienting axes* (eg., front/back, left/right, see figure 3). A generating axis is central to Marr's accounts of object shape and determines the object's principal axis (cf. [Marr 82]). For instance, the generating axis of a human body is vertically oriented. In our model, the orienting axes are orthogonal to the generating axis and to one another (e.g., corresponding to the front/back and side/side axes; for more details cf. [Landau & Jackendoff 93]).

Recent studies support the hypothesis that a nonlinguistic disparity between the representation of *what* and *where* underlies how language represents objetcs and places (cf. [Landau & Jackendoff Kosslyn asserts that the spatial component of mental imagery preserves *where* information about the meaningful parts of a scene and their relative locations, whereas the visual component preserves *what* information about how an image looks (cf. [Kosslyn 87]). In this sense, the object representation of a landmark is used to determine *what* information, whereas the generalized representation serves for *where* information. Thus, geometrical or physical and

conceptual information about landmarks refer to the object representation and the determination of spatial relations between landmarks to the generalized representation.

In our model, we only use some restricted salience criteria: size, shape, and colour. The visual perception process identifies landmarks without request (automatic visual perception) or by demanded of the planning process (controlled visual perception). The planning process uses domain-specific strategies in order to request for specific landmarks. These strategies provide alternative objects to be looked for if the visual perception process cannot find an requested object. Thus, the visual perception process can be described as follows:

1. While there is no request, identify objects with the strongest saliency (automatic visual perception)
2. If there is a request for a specific object look for it in a default region (controlled visual perception)
3. If this specific object cannot be found report it to the planning process and goto step 1.

An example for automatic visual perception is the perception of a red telephone box in a German city. We attend this telephone box although we might not look for it. An example for controlled visual perception is when we drive by car towards a crossing. Then the following strategy is normally used: First look for a traffic light somewhere in the upper right visual area. If the visual perception process cannot find a traffic light it looks for a traffic sign. If there is no traffic sign, supply the most salient object.

3.2 How is map knowledge used for wayfinding?

Visual perception provides visual reconstruction access to the immediate environment. To find a non-trivial path, we commonly use maps to orient ourselves and to determine an optimal path considering several constraints. When humans use a map, they extract an abstract path. This is mainly determined by a direct connection line between the starting point and the destination and by a search corridor around this path. This area, which we call the *rough path*, is determined before the speaker `SP` starts navigating and is elaborated stepwise while moving towards the destination. Incremental use of the map appears to be due to the limited capacity of working memory. We have developed a waysearch process

that is based on this type of incremental search. In order to get some data on how people use maps we have conducted several interviews.

In work done by Klein, the planning of the primary plan for complete route descriptions is divided into two techniques: advanced and stepwise planning. Klein proposed that these techniques can be used in combination. Advanced planning means that the route description process is mostly sequential, i.e. that first the wayfinding process determines the whole path. Then, this path is described without going back to the wayfinding process. In contrast to this, stepwise route planning implies that there is an interaction between the wayfinding and the route presentation process (cf. [Klein 82]). Gluck remarks that the potential optimizing functions that are relevant for the wayfinding process are not restricted to minimal distance or even minimal effort (cf. [Gluck 91]). The wayfinding process in the TOUR model is a simple and slow approach to finding a way (cf. [Kuipers 77]). In the TRAVELLER model proposed by Leiser and Zilbershatz an incomplete wayfinding process is used which cannot find partial paths in the cognitive map (cf. [Leiser & Zilbershatz 89]).

We distinguish between the selection of a path that is known by experience (experience-based wayselection) and the search for a path by using a map (cf. [Elliott & Lesk 82]). In the TOUR model (see cf. [Kuipers 77]) and in the TRAVELLER model (cf. [Leiser & Zilbershatz 89]) experience-based approach is focussed. Both models are based on a graph network which represents a 'mental' model of the spatial environment and concentrate on how to integrate new knowledge about the *external* world into it. Leiser and Zilbershatz divide route knowledge into a *basic network* for the main routes and a *secondary network* for the other routes. This division is based on experiments done by Pailhous (cf. [Pailhous 69]) and Chase (cf. [Chase 82]). A similar approach is presented by Gopla et al. by their NAVIGATOR model (cf. [Gopal et al. 89]). In this model they use a hierarchical structure for representing knowledge about the spatial environment.

The waysearching process is not as well understood as the experience-based wayselection process. In most of the relevant work, the route that has to be presented is always known or generated by some straightforward algorithm like the A*-algorithm[3]. The waysearch process has also to observe various kinds of constraints. If the person goes by car, the search process has to consider streets accessible by cars and the type of street. For instance, for long-distance trips car-driver prefer to use highways and no side-paths. It is also relevant to consider the weather, e.g. if it rains you will prefer to describe a covered path to a walker.

[3]It is very unlikely that common heuristic search algorithms, like A*, are similiar to human search ability in maps.

There is some evidence that the cognitive ability of wayfinding combines the experience-based wayselection process and the waysearch process. For example, it is not usual that we remember a large scale space route without searching for some path segments. Combining wayselection and waysearch is what we do when we navigate along a route.

When we asked people to find a non-trivial path they tended to use a restriced area in which they try to determine path segments. This area was mainly determined by a imaginary line between the starting point and the destination and by an area around this line. People who were not familiar with the given map determined the path stepwise.

Basing on these results, we have developed an incremental search algorithm which integrates most of the above mentioned constraints. This algorithm is a modified version of the RTA*-algorithm (cf. [Korf 90]). One of the main features of this algorithm is that search is directed by a spatial corridor, which serves as a model for a rough path. The search algorithm prefers paths inside the corridor. Paths outside of the corridor are considered but weighted according to their distance from the corridor. The algorithm computes the next relevant path segment from the current turning-point to the next. Thus it transforms the rough path into topographical path segments.

3.3 Integration of map knowledge and the knowledge given by visual perception

Information about landmarks and path segments given by visual perception has to be interrelated with topographical information obtained from a map. In each situation the speaker has to combine four different kinds of information to describe a route while navigating along the path. First, the rough path is used to keep track of global orientation and it restricts the search area in a map. With the help of a rough path a local topographical part of the path can be selected from a map. After that, the speaker has to identify the path segment in the current environment. Representation of landmarks obtained by visual perception are the third kind of information. On the fourth level, those spatial relations are noted which habe been established by `SP` for relevant landmarks perceived from `SP`'s position (cf. [Gapp 94]). All three kinds of information are integrated in a central structure, called a *segment* which serves as the internal spatial representation used for route descriptions.

Consider for instance a path segment where the speaker stands at position `A` and is oriented towards point `B` (see figure 4). The topographical path segment between point `A` and `B` is a projection of the rough path on the map. In this case A and B are turning-

points. Associated with this path segment are landmarks `tower-1` and `house-4` determined by salience criteria.

A segment guides the speaker's orientation, navigation and route description for a specific time interval. The length of this interval depends on the underlying spatial structure and the kind of navigation. If a non-trivial route has to be described, the speaker does not keep a description of the whole route readily mentally accessible but only the part which is currently relevant. Segments are mainly determined by large changes in orientation. For instance, if the speaker `SP` turns left at a crossing `SP` usually enters a new spatial environment which was not visible before the turn. Information necessary before `SP` turned left is in general useless after the turn, but it is still accessible and `SP` can refer to it which demands some kind of history. Each segment is added to a general spatial representation format, called the *segmentation structure*. The segmentation structure is the whole spatial representation from the starting point to the current situation seen from the speaker's perspective.

A segment is the basis for the presentation planning process. We will present an example to clarify what is meant by a segment and by the segmentation structure. Suppose the spatial configuration shown in figure 4. The goal of `SP` is to present a route from the *starting point* `S` to the destination point `D`. On the choosen route there are three turning-point landmarks (`A`, `B`, and `C`). Between `S` and `A` is one landmark, a tower on the right side, and between `A` and `B` are two landmarks, namely a house on the right side and a bridge over the street. The streets are labelled by *street-1* to *street-4* started at `S`.

The topographical path level of a segment represents path segments between the starting point `S`, the turning-points `A`, `B`,`C`, and the destination `D`. Empirical data show that turning-points are frequently used landmarks in route descriptions (cf. [Klein 82, Wunderlich & Reinelt 82, Habel 87, Meier et al. 88, Hoeppner et al. 90]).

On the third level, all landmarks accessed by visual perception are integrated, like houses, trees, human beings, and animals. The difference between the second and the third level is that we switch from *large-scale space* to *small-scale space*. "Large scale space is space whose structure cannot be observed from a single viewpoint ([Kuipers 78])". In contrast to this, small-scale space includes all spatial objects which can be seen from a single viewpoint. An example for use of large-scale space information in route descriptions is that it is not necessary to perceive the next turning-point from the current position. Small-scale space information is needed when we mention a house on the right side. Then it is necessary that the house is currently visible.

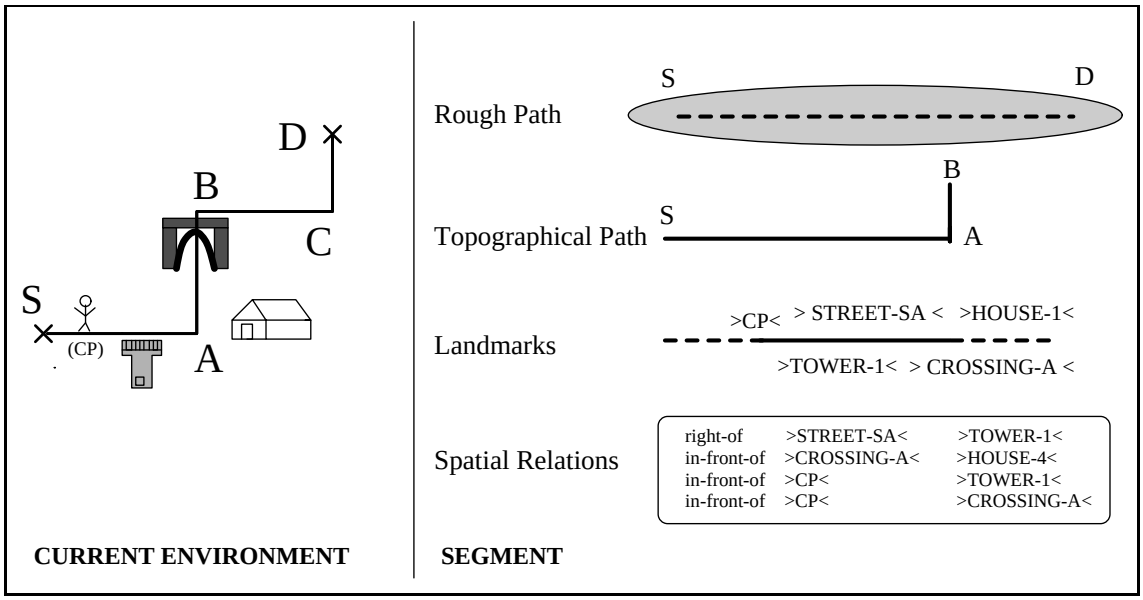

Figure 4: Complete segmentation structure: How to come from an environment to a segmentation structure.

Every time when SP leaves the current path segment SP has to determine the next one. For example, if SP turns left at A, SP leaves SA and reaches AB that becomes the current path segment. Now, SP has to determine path segment BC to be able to describe the transition from segment AB to BC.

The whole segmentation structure is incrementally generated on the way from the starting point to the destination. This reflects the observation that SP mostly does not know the whole topograhical path in detail at the beginning. Another observation is that for humans it is easy to choose a new path if they have gone the wrong way. In the proposed model a new rough path, the new segment and its successor have to be determined to proceed. Starting from a given segment, the planning process determines a central presentation structure. Here, the spatial representation and the linguistic representation are integrated, similar to the Landau and Jackendoff's approach (cf. [Landau & Jackendoff 93]). Those parts of the segmentation structure which are intended to be integrated in the communication are selected by the planner.

4 Planning process

The integration of spatial knowledge and path information in order to use it in natural language descriptions requires flexible control processes. Therefore we decided to use a plan-based approach. A central planning problem concerns the handling of temporal aspects of both the generation and following of route descriptions; this problem has received relatively little attention in previous planning research. Temporal relations, e.g., "Turn left *immediately/after 5 minutes*", are an important part of incremental route descriptions. Spatial relations involved in imperatives of movement ("Keep going along the field!") only hold for a certain amount of time. Thus, the planner has to integrate both spatial and temporal information. Beside these temporal relations concerning the content of the information, the planning process has also to consider temporal constraints given by relations between speaker, hearer, and the current environment, e.g., time the speaker and hearer need until they reach the crossing. Also, different strategies have to be compared to apply the best strategy in a given situation.

4.1 How do we plan the presentation?

In *complete route description*, there are two competing strategies: planning in advance and stepwise planning (cf. [Klein 82, p.7]). Because of the incrementality, the process can be characterized as a stepwise planning process. At the starting point there is only a *rough path* to the destination point and the information for the first segment. The main task of the planning process is the conversion of spatial information into presentation-oriented information, but it also determines the temporal structure of the presentation.

In our approach, the planning process deals with two major problems: Landmark selection and establishment of spatial relations. Salience of landmarks depend on various parameters, e.g. form, colour, and function of a landmark, light and weather conditions etc. (see section 3.1). We define a *spatial relation* by specifying certain conditions concerning a landmark in a spatial configuration. Spatial relations are considered to be independent of any particular natural language and they form the basis for the analysis of spatial prepositions within the framework of reference semantics (cf. [Schirra 90, Gapp 94]).

Landmarks selected by visual perception are not isolated from each other, but can be interrelated by spatial relations. Following Landau and Jackendoff, spatial relations mainly depend on boundedness, surface, or volumetric nature of a landmark and its axial structure

VERBAL(	M-TYPE	by-feet
	TIME	(Δt_1)
	ACTUAL	(type(street-1))
	UNTIL	(type(crossing-1))
	ACTION	(move-straight))
VERBAL(	M-TYPE	(by-feet)
	TIME	(Δt_2)
	LM	(type (house-4, attr(colour, grey)))
	SPATIAL-REL	(in-front-of,turn-point-1,house-4)
	SPATIAL-REL	(left-of, street-2, street-1)
	ACTION	(turn-left))

Figure 5: Two presentation stuctures for natural language generation

(cf. [Landau & Jackendoff 93, p.232]). For instance, in figure 4 the landmarks `bridge-1`, `house-4`, and `tower-1` can be described by object features or by their spatial relationship with one another.

The transformation of information extracted from a segment into a presentation structure is done by two techniques: schematic and non-schematic (cf. [Maaß 93]). It is clear that the usage of schemata depends on how familiar we are with the current environment and the intended description. Beside dependencies concerning the hearer and environment, external parameters, like the time of day, are important. For example, a landmark may be relevant in the night because it is illuminated, but not during the day.

In general, there are no problems to switch between the presentation modes if we want to inform the questioner about, for example, a red house on the right side of a street. In `MOSES`, we do this with a verbal expression, with a pointing gesture or a line drawing. Therefore we propose a central presentation representation. When the planning process selects items of a segment in order to integrate it into the presentation representation, it has to consider the cross-referentiality of information (cf. [Wahlster et al. 92]). This means that an item can be parallely presented in different modes. To solve this problem the planning process keeps track of the valid use of information in different presentation modes.

The planning process distributes information for the presentation representation according to the individual strength of each presentation mode. For presentation processes we use approaches (cf. [Herzog et al. 93, Maaß 93]). For explanation of the main points, we give a simple example where the turn from path segment `SA` into `AB` is represented by two presentation structures (see figure 4.1):

The modespecific presentation structure (cf. figure 4.1) provides global information, like the presentation mode and the way the questioner moves (M-TYPE). By using this and the action type (ACTION), a movement verb is determined. It also contains information about the current path segment, ACTUAL(type(street-1)), and the following path segment, NEXT(type(street-2)). Another important item of the structure is the next turning-point, UNTIL(type(crossing-1)), that defines the connection to the next path segment. The time interval t_i describes the slot when the information is intended to be presented. All landmarks selected by the presentation planner are gathered in the landmark list (LM). Those spatial relations between these landmarks and the speaker which are to be described are given by a list (SPATIAL-REL).

With this presentation structure the speech process generates the following textural description:

Go down the street, until you reach the next crossing.	(during Δt_1)
In front of the grey coloured house turn left.	(during Δt_2)

4.2 Modespecific Presentation Processes

In MOSES, we use several presentation modes: 2-dimensional graphical projection, perspective graphical animation, incremental natural language, and pointing gestures. The environment is represented by a 3-dimensional geometric model (see figure 6).

Survey knowledge about the region in question is represented by a 2-dimensional projection onto the horizontal plane. This presentation mode alone is not optimal because of the cognitive effort the user has to spend matching the projected landmarks or crossroads with their actual visual appearance. Here, perspective views that simulate the speaker's perspective can help. Both graphical presentation modes can be supplemented with several display techniques, such as highlighting, flashing, etc.

In MOSES, the incremental generation of surface structures is realized with an incremental generator for German and English (cf. [Finkler & Schauder 92]), which is based on *Tree Adjoining Grammars* (cf. [Joshi 83]). For spontaneous gestures, we have developed a prototypical module for generating arrow-based pointing gestures.

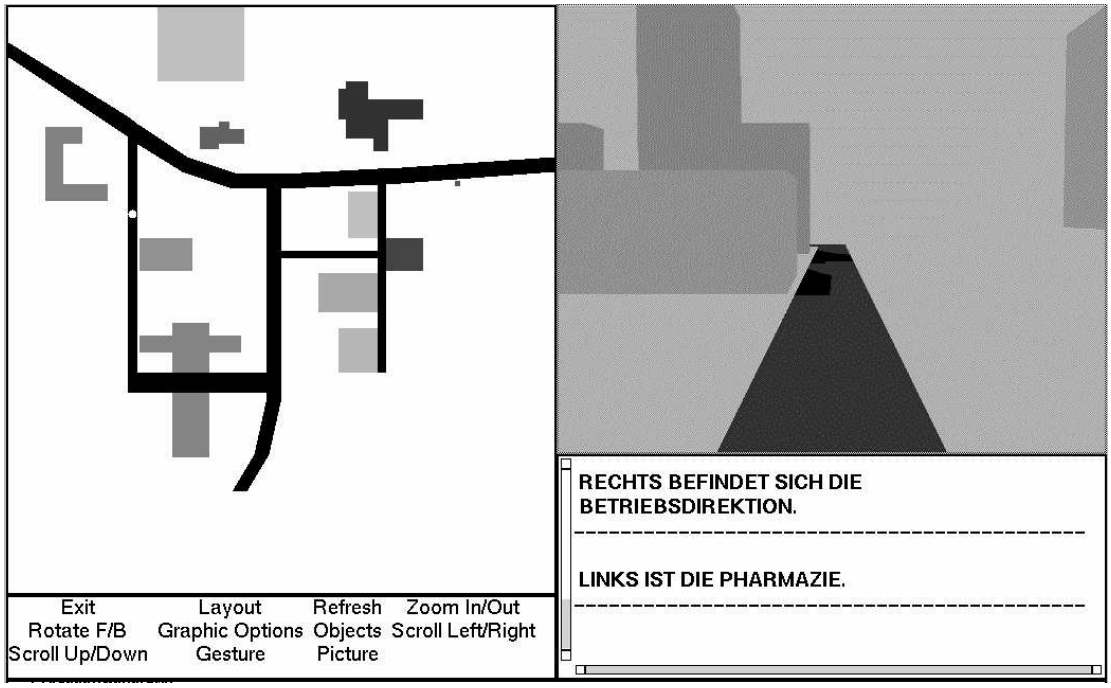

Figure 6: A view on MOSES

5 Conclusion and open problems?

Research on route descriptions guided by visual perception leads to the conclusion that we have to assume at least four subprocesses: a visual perception process, a wayfinding process, a presentation process, and a planning process. Central for the representation of spatial information is the incrementally generated segmentation structure. The interaction between spatial information and presentation is done by the planning process. It enables convergence of information from the segmentation structure by selecting relevant information and by establishing spatial relations between landmarks and the speaker's position. Information about landmarks is obtained from a dualistic object representation.

We presented a global framework, that has to be refined, perhaps modified, and compared with results of psychological research in the future. Especially the interaction between the subprocesses and the representation structure has to be examined in more detail. Our current work is concerned with the seamless integration of temporal constraints. We also evaluate wether the spatial representation and the presentation representation are sufficient for route descriptions in dynamic scenes.

References

[Allen & Kautz 85] J. F. **Allen** and H. A. **Kautz**. *A Model of Naive Temporal Reasoning.* In: J. R. Hobbs and R. C. Moore (eds.), Formal Theories of the Commensense World, pp. 251–267. Norwood, NJ: Ablex, 1985.

[Biederman 90] I. **Biederman**. *Higher-Level Vision.* In: D. N. Osherson, S. M. Kosslyn, and J. M. Hollerbach (eds.), Visual Cognition and Action: An Invitation to Cognitive Science (Volume 2), pp. 41–72. Cambridge, MA: MIT Press, 1990.

[Blades 91] M. **Blades**. *Wayfinding Theory and Research: The Need for a New Approach.* In: D. M. Mark and A. U. Frank (eds.), Cognitive and Linguistic Aspects of Geographic Space, pp. 137–165. Dordrecht: Kluwer, 1991.

[Chase 82] W.G. **Chase**. *Spatial Representations of Taxi Drivers.* In: D.R. Rogers and J.A. Sloboda (eds.), Acquisition of Symbolic Skills. New York: Plenum, 1982.

[Elliott & Lesk 82] R. J. **Elliott** and M. E. **Lesk**. *Route Finding in Street Maps by Computers and People.* In: Proc. of AAAI-82, pp. 258–261, Pittsburgh, PA, 1982.

[Finkler & Schauder 92] W. **Finkler** and A. **Schauder**. *Effects of Incremental Output on Incremental Natural Language Generation.* In: Proc. of the 10 th ECAI, pp. 505–507, Vienna, 1992.

[Gapp 94] K.-P. **Gapp**. *On the Basic Meanings of Spatial Relations: Computation and Evaluation in 3D Space.* In: To appear in: Proc. of the 12 th AAAI-94, Seattle, WA, 1994.

[Garling 89] T. **Garling**. *The role of cognitive maps in spatial decisions.* Journal of Environmental Psychology, 9:269–278, 1989.

[Glasgow 93] J. **Glasgow**. *The Imagery Debate revisited: A Computational Perspective.* Computational Intelligence, 9(4):309–333, 1993.

[Gluck 91] M. **Gluck**. *Making sense of Human Wayfinding: Review of Cognitive and Linguistic Knowledge for Personal Navigation with a new Research Direction.* In: D.M. Mark and A.U. Frank (eds.), Cognitive and Linguistic Aspects of Geographic Space, pp. 117–135. Netherlands: Kluwer Academic Publishers, 1991.

[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.

[Grosz 81] B. J. **Grosz**. *Focusing and Description in Natural Language Dialogues.* In: A. Joshi, B. L. Webber, and I. A. Sag (eds.), Elements of Discourse Understanding, pp. 84–105. Cambridge, London: Cambridge University Press, 1981.

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

[Hayes-Roth & Hayes-Roth 79] B. **Hayes-Roth** and F. **Hayes-Roth**. *A cognitive model of planning.* Cognitive Science, 3:275–310, 1979.

[Herzog et al. 93] G. **Herzog**, W. **Maaß**, and P. **Wazinski**. *VITRA GUIDE: Utilisation du Langage Naturel et de Représentation Graphiques pour la Description d'Iténitaires.* In: Colloque Interdisciplinaire du Commitée National "Images et Langages: Multimodalitée et Modélisation Cognitive", Paris, 1993.

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

[Ittelson 60] W. H. **Ittelson**. *Visual Space Perception.* New York: Springer, 1960.

[Jackendorf 87] R. **Jackendorf**. *On Beyond Zebra: The Relation of linguistic and visual Information.* Cognition, 26:89–114, 1987.

[Joshi 83] A. K. **Joshi**. *Factoring Recursion and Dependencies: An Aspect of Tree Adjoining Grammars (TAG) and a Comparison of Some Formal Properties of TAGS, GPSGS, PLGS, and LPGS.* In: Proc. of the 21 th ACL, pp. 7–15, Cambridge, MA, 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.

[Korf 90] R. E. **Korf**. *Real-Time Heuristic Search.* Artificial Intelligence, 42:189–211, 1990.

[Kosslyn 87] S. **Kosslyn**. *Seeing and imaging in the cerebral hemispheres: a computational approach.* Psychological Review, 94:148–175, 1987.

[Kuipers 77] B. **Kuipers**. *Representing Knowledge of Large-Scale Space.* PhD thesis, MIT AI Lab, Cambridge, MA, 1977. TR-418.

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

[Landau & Jackendoff 93] B. **Landau** and R. **Jackendoff**. *"What" and "where" in spatial language and spatial cognition.* Behavioral and Brain Sciences, 16:217–265, 1993.

[Leiser & Zilbershatz 89] D. **Leiser** and A. **Zilbershatz**. *THE TRAVELLER: A Computational Model of Spatial Network Learning.* Environmental and Behaviour, 21(4):435–463, 1989.

[Lewis 76] D. **Lewis**. *Observations on Route Finding and Spatial Orientation among the Aboriginal Peoples of the Western Desert Region of Central Australia.* Oceania, XLVI(4), June 1976.

[Maaß et al. 94] W. **Maaß**, J. **Baus**, and J. **Paul**. *Visually Accessed Spatial Information Used in Incremental Route Descriptions: A Computational Model.* In: submitted, 1994.

[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.

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

[McCalla & Schneider 79] G. **McCalla** and P. **Schneider**. *The Execution of Plans in an Independent Dynamic Microworld.* Proc. of the 6 th IJCAI, pp. 553–555, 1979.

[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.

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

[Pailhous 69] J. **Pailhous**. *Representation de l'espace urbain et cheminements.* Le Travail Humain, pp. 32–87, 1969.

[Piaget et al. 60] J. **Piaget**, B. **Inhelder**, and A. **Szemenska**. *The Child's Conception of Geometry.* New York: Basic Books, 1960.

[Schirra 90] J. R. J. **Schirra**. *Einige Überlegungen zu Bildvorstellungen in kognitiven Systemen.* In: C. Freksa und C. Habel (Hrsg.), Repräsentation und Verarbeitung räumlichen Wissens, pp. 68–82. Berlin, Heidelberg: Springer, 1990.

[Schneider & Shiffrin 77] W. **Schneider** and R. M. **Shiffrin**. *Controlled and automatic human information processing. 1. Detection, search, and attention.* Psychological Review, 84:1–66, 1977.

[Siegel & White 75] A. W. **Siegel** and S. H. **White**. *The Development of Spatial Representation of Large-Scale Environments.* In: W. Reese (ed.), Advances in Child Developement and Behaviour. New York: Academic Press, 1975.

[Sitaro 93] K. **Sitaro**. *Language and Thought Interface: A Study of Spontaneuos Gestures and Japanese Mimetics.* PhD thesis, University of Chicago, August 1993.

[Stokols & Altman 87] D. **Stokols** and I. **Altman** (eds.). *Handbook of Environmental Psychology*, volume 1 & 2. John Wiley & Sons, 1987.

[Streeter et al. 85] L.A. **Streeter**, D. **Vitello**, and S.A. **Wonsiewicz**. *How to Tell People Where to Go: Comparing Navigational Aids.* International Journal of Man-Machine Studies, 22:549–562, 1985.

[Thorndyke & Goldin 83] P. W. **Thorndyke** and S. E. **Goldin**. *Spatial Learning and Reasoning Skill.* In: H. L. Pick and L. P. Acredolo (eds.), Spatial Orientation: Theory, Research, and Application, pp. 195–217. New York, London: Plenum, 1983.

[Thorndyke & Hayes-Roth 82] P.W. **Thorndyke** and B. **Hayes-Roth**. *Differences in Spatial Knowledge Acquired from Maps and Navigation.* Cognitive Psychology, 14:560–582, 1982.

[Tolman 48] E.C. **Tolman**. *Cognitive Maps in Rats and Men.* Psychological Review, 55:189–208, 1948.

[Wahlster et al. 92] W. **Wahlster**, E. **Andre**, W. **Finkler**, W. **Graf**, H.-J. **Profitlich**, T. **Rist**, and A. **Schauder**. *WIP: Integrating Text and Graphics Design for Adaptive Infromation Presentation.* In: R. Dale, E. Hovy, D. Rösner, and O. Stock (eds.), Aspects of Automated Natural Language Generation: Proc. of the 6th. International Workshop on Natural Language Generation, pp. 290–292. Berlin, Heidelberg: Springer, 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.

[Yeap 88] W. K. **Yeap**. *Towards a Computational Theory of Cognitive Maps.* Artificial Intelligence, 34:297–360, 1988.