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. 102

Spatial Layout Identification and Incremental Descriptions

Klaus-Peter Gapp, Wolfgang Maaß

Mai 1994

ISSN 0944-7814

102

Spatial Layout Identification and Incremental Descriptions

Klaus-Peter Gapp and Wolfgang Maaß
Cognitive Science Program
University of Saarbrücken
66041 Saarbrücken, Germany
Email addresses: {gapp, maass}@cs.uni-sb.de

Abstract

The integration of perception sources is one of the most interesting topics in AI. In this context, the question: How can we discuss what we see? is especially prominent. Nevertheless, there is very little research activity in this direction. In the project VITRA, we are concerned with the interaction of high-level visual perception and speech production. Using the domain of route descriptions, we present how to determine and how to use spatial information of the current static 3-dimensional environment in verbal utterances. First, we show how layout information, i.e., object information and spatial relations, is determined in the current environment. In the second part, we demonstrate the use of layout information in verbal movement descriptions.

1 Introduction

Although the understanding of vision and natural language production are main research topics in AI, there has been very little attention in integrating both fields. This stands in contrast to related investigations in psychology, especially forced by the imagery debate (cf. [Pinker 84]). Also neurological research of the brain presents findings in the neural connection between spatial/visual representations and language (cf. [Kosslyn et al. 90]). When we talk about what we see, we must take into account how we experience and behave in our environment. Beside psychologists and philosophers, also geographers and especially urban planners ([Lynch 60]) are interested in how people learn and talk about the world in everyday life.

Here, we discuss how we describe travel actions in static 3-dimensional environments. Therefore, we use the area of *incremental route descriptions*, where a speaker P and a hearer H travel along a path and P verbally describes H where to go step-by-step, like a co-driver. We present how visual perception and speech production can be treated in a common framework. Specifically, we focus on the computation of object information, spatial relations, and the use of both for incremental verbal descriptions. In this context two global questions arise: (1) How to identify the layout in a scene? and (2) How to describe the current environment by using layout information? By layout we mean the location of objects and their spatial relations to one another in the current environment as seen from P's perspective (deictic use).

The phenomena of route learning and navigation are simulated by several models (eg., [Kuipers 78; Gopal et al. 89]) but none of these are concerned with the integration of visual perception and speech generation. We have developed a model for the process of *mul-*

timodal, incremental route description, called MOSES ([Maaß 94]). The model consists of a high-level visual perception[1] process, a way-finding process, a planning process, and two presentation processes: speech generation and gesture generation (cf. [Maaß 94]).

The way-finding process models the human ability to choose a route from a map in combination with object information obtained by visual perception. We use a dualistic representation. On one hand, the speaker has an overview which includes orientation and distance information. On the other hand, s/he uses this general information to produce an exact topographical partition of the path which leads his/her movement. This general information is also used to produce the description for the current spatiotemporal interval (for more details see [Maaß 94]). Interactions between visual perception, way-finding, and presentation processes are supervised and coordinated by a multi-agent planning process.

First, we shortly discuss our approach to visual perception and in specific object representation (section 2). In section 3 we present how to compute spatial relations between identified objects using a multilevel semantic model (cf. [Gapp 94b]). We agree with [Landau & Jackendoff 93] thesis that if people are applying spatial relations they do not take into account every detail of the objects involved. We are therefore able to use an *approximative* algorithm, which considers only some necessary shape properties of an object. In section 4, we show how layout information is used to determine natural language utterances concerning incremental route descriptions.

This work is part of the project VITRA (VIsual TRAnslator) which deals with the relationship between natural language and vision (cf. [André et al. 89]).

[1]Following David Marr's representational framework for vision ([Marr 82]), we associate low-level vision with processes and representations concerned with what Marr calls the *primal sketch* and the *2-1/2 sketch* and high-level vision with the *3D model*.

2 Visual Perception

In MOSES, we do not deal with low-level aspects of visual perception (cf. [Marr 82]) but we start at an object level. In the current visual field we focus on an *inner* visual field, in which we can identify, categorize and select distinct objects. Hence, the inner visual field attracts all attention of the visual perception process. The current inner visual field determines the objects which can be used in the descriptions: Visual perception leads communication. If the visually accessed information is not sufficient, visual perception will be led by the demands of the presentation. What we want to know is: How can high-level visual data be processed efficiently for use in communication?

In a given situation, the environment provides a vast amount of information. But only a part of it is selected by the speaker P for inclusion in the utterance. In our approach, we only use a restricted set of salience criteria: size, colour, and brightness. Visually selected objects are represented by 3D models (cf. [Maaß 94]). In MOSES, the 3D model representation is used as an interface between vision and natural language processing (see also [Landau & Jackendoff 93]). Objects selected by visual perception are not isolated from each other, but can be interrelated by spatial relations.

In [Herskovits 86, p. 57ff] Herskovits proposed employing different kinds of object idealizations, e.g., object approximations to a point, a line, a surface, a horizontal plane, etc. Landau and Jackendoff also confirm that spatial relations mainly depend on boundedness, surface, or volumetric nature of an object and its axial structure [Landau & Jackendoff 93, p. 226]: *"What proves surprising is how sparsely both the figure and reference objects appear to be represented. ..., there are no prepositions that insist on analysis of the figure or reference object into its constituent parts."*.

Therefore it seems to be reasonable to consider only approximated shape properties of an

objects when computing spatial relations. In most cases, it is sufficient to approximate the object to be localized with its center of gravity, because its position is the only necessary information that counts for the applicability of the spatial relation. In our system we are using the following simplifications at the moment:

(1) *Center of gravity*. (2) *2D representation*: The base of each object (Necessary when perceiving objects from a bird eye's view, e.g., in maps). (3) *Bounding rectangle* (*BR*): The bounding rectangle of an object with respect to a direction vector $\vec{v}$ is the minimal rectangle which is aligned to $\vec{v}$ and contains the 2D representation of the object (Figure 1a). (4) *Bounding right parallelepiped* (*BRP*): The bounding right parallelepiped of an object with respect to a direction vector $\vec{v}$ is the minimal right parallelepiped which is aligned to $\vec{v}$ and contains the object (Figure 1b). (5) *3D representation*: The complete description of an object.

Beside a *geometric representation*, the representation of an object integrates a *conceptual representation* which includes size, colour, and intrinsic properties of the object (cf. [Gapp 94b]).

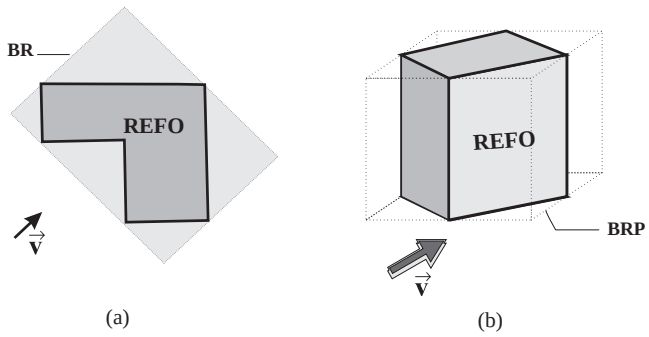

Figure 1: *BR* and *BRP*

3 Spatial relations

Spatial relations play a key role in the process of connecting visual and verbal space. They act as a connecting link between visually perceived data and natural language. Spatial relations are categorized in two main classes: the topological and the projective.[2] The linguistic representatives for the spatial relations are the prepositions in their spatial meanings. Combined with a statement of place they build the class of *localization expressions* [Herskovits 86].

A purely geometrical representation of the semantics of spatial relations is not appropriate (cf. [Miller & Johnson-Laird 76]), because functional dependencies or pragmatic principles, as Grice's principles of cooperativity [Grice 75], are not considered. Therefore we propose the use of a multilevel model [Gapp 94b], influenced by the existing 2-level [Bierwisch & Lang 89] and 3-level [Aurnague et al. 90] approaches. Our multilevel model allows us to take into account just the geometrical shape properties of the objects (*basic meanings*) and enables us to include functional dependencies and pragmatic issues on a higher level.

As mentioned before, people do not account for every detail of the objects involved when applying spatial relations. We are therefore able to use an *approximative* algorithm, which considers only the necessary shape properties of an object for the computation, e.g., the center of gravity or the smallest circumscribing rectangle.

The algorithm for the computation of the applicability of a topological relation's basic meaning considers only the distance between the object to be localized (*LO*) and the reference object (*REFO*) scaled by the *REFO's* extension in each of the three dimensions (*local distance*).

In addition to the computation of projective relations we need to include the scaled deviation angle of the from the canonical direction implied by the relation (*local angle*). The dependency of the local distance and the local angle from the REFO extension ensures for a bigger *REFO* in a dynamic enlargement of a relation's region of applicability (cf. [Gapp 94a]). This evaluation

[2] And additionally the relation *between*, which needs two reference objects and thus takes an exceptional position in the group of spatial relations.

process can be applied to 2-dimensional as well as to 3-dimensional environments. In the latter case, we get a 3-dimensional extensive region of application for spatial relations such as *at* or *in front of*. This seems reasonable when we look at Figure 4b, because *"turn right in front of* L_4*"* is an adequate application of the spatial relation *in front of* although there is a vertical difference between the house L_4 and the street segment S_4.

The evaluation of the applicability of a spatial relation differs from one person to another [Kochen 74]. Because of this we use user adjustable cubic spline functions, which enable a cognitively plausible continuous gradation of the relation's applicability. To get an idea of a region's applicability structure we extend the so-called *potential fields* presentation of a spatial relation (cf. [Yamada et al. 88; Schirra 90]) to the third dimension. For instance, Figure 2 shows cross-sections of the 3-dimensional potential field of the relation *right* concerning a high building as reference object. Its large extension in the vertical is reflected in the shape of its applicability region.

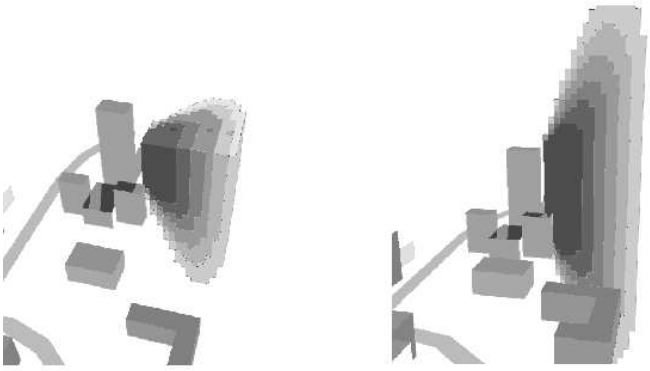

Figure 2: Cross sections of the potential field: *"right of the high building"*

Adding a certain context involves mostly modified regions of applicability for the basic meanings. For example, in Figure the landmark[3] L_4 is bigger than the landmark L_1 and therefore — concerning the basic meanings — the region of applicability $RA^{vor}_{L_4}$ of the relation *in front of* with respect to L_4 as the *REFO* is bigger than $RA^{vor}_{L_1}$ (Figure 3a,b).[4] But if a street runs between them, the shape of the two regions changes (Figure 3c).

We end this section with a brief discussion of how perspective view might influence the application of spatial relations. The main influence arises when reference objects are only partially visible. Although we are not able to perceive the whole shape of the *REFO* we are usually able to establish spatial relations referring to it. For instance in Figure 4b the house L_4 is partially hidden. Nevertheless, the direction *"turn right in front of* L_4*"* is clearly understood.[5] Apparently that we assign some kind of defaults to the extension of the hidden reference object and these default values are used to compute the spatial relations. In addition world knowledge helps us in building relationships. In Figure 4b it is obvious for us that the street segment S_4 (cf. Figure 4a) leads away perpendicular to S_1, and therefore must be in front of L_4.

4 Incremental route descriptions

We now discuss how layout information is used to adequately describe a given scene while following a path. Two major problems can be identified while determining an adequate incremental route description: First, a selection problem and second a temporal problem. By selection problem, we mean that not all information obtained by the environment is used in a description. Most of the information is not relevant for the hearer. On the other hand, the speaker must determine the correct temporal descriptions (cf. [Maaß 94]).

[3] With landmark we denote objects which are salient to P in the current environment.

[4] The region of applicability are drawn without gradation for reasons of a better readability.

[5] For readability, we use english terms for all spatial expressions although we have examined German descriptions.

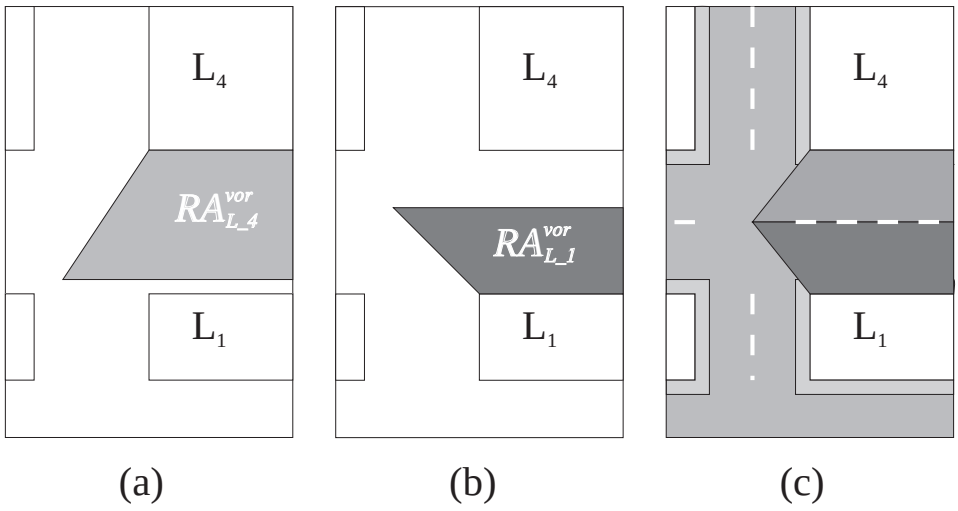

Figure 3: Influence of functional dependencies

The information flow from visual perception to speech production is divided into two subsystems: *Input system* and *output system*. In the input system, information about landmarks and path segments selected visually must be interrelated with topological information obtained from a map. Representations of landmarks observed and spatial relations between these landmarks are integrated in a central structure, called a *segment* which serves as the internal spatial representation of the current layout (cf. [Maaß 93; Maaß 94]).

spatial relation	LO	REFO	degree of applicability
right-of	S_1	L_2	0.8
left-of	S_1	L_1	0.8
behind	C_1	S_1	1
left-of	C_1	S_4	1
before	L_1	C_1	0.7
behind	L_4	C_1	0.7
in-front-of	S_4	L_4	0.9
behind	S_4	L_1	0.8
on	P	S_1	1

Tabular 1: Extracted spatial relations

Up to now, we have demonstrated how spatial information is obtained from the environment and represented as internal structures. In order to describe the current environment adequately, the speaker must select the information which seems to be relevant to the hearer. Here, we do not address the problem of anticipating the hearers beliefs; instead of a hearer model we use default assumptions.

We will now outline the key ideas of our approach by using an example situation (see figure 4). Suppose that the speaker stands at position `P` marked by a cross and `P's` intention is to turn right at the crossing C_1 into street segment S_4. `P` has visually obtained four street segments (S_i), four landmarks (L_i), and a crossing (C_i). We assume that `P` has obtained layout information from the environment and `P` has integrated this information into the current segment. Because we assume that visual perception and speech production are independent from each other, more information about the environment can be integrated into the current segment step-by-step as `P` travels along the path. In contrast to this, we assume that visual perception and way-finding are exclusive tasks. This means that `P` can either look at the current environment or on the map. In addition to object information, `P` determines spatial relations between these objects (S_i, L_i, and C_i) of the current layout (see tabular 1).

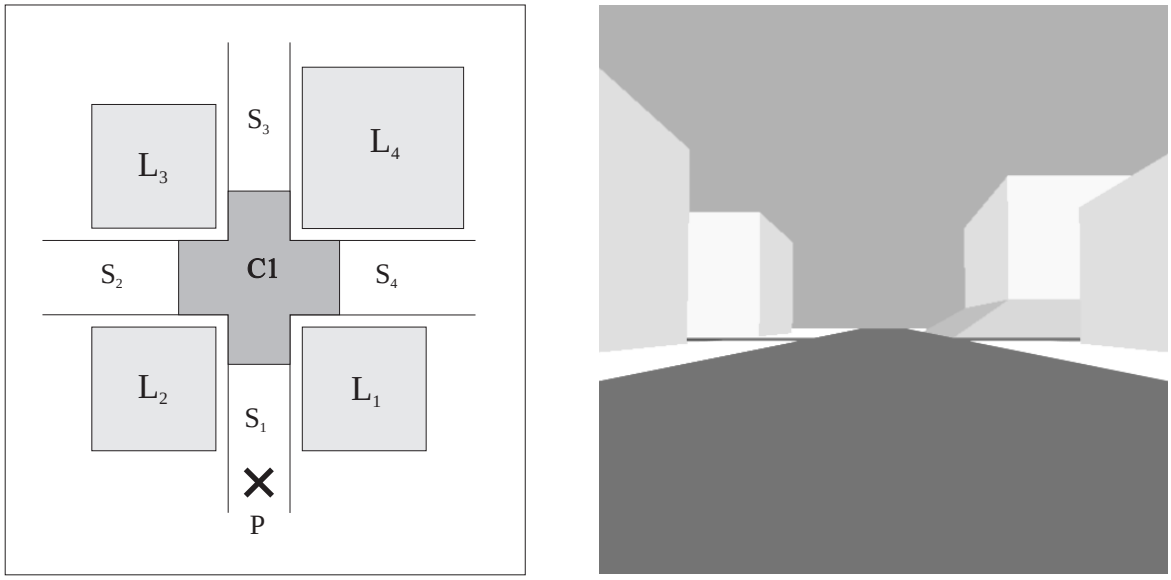

Figure 4: Example layout

We have examined several transcripts of route descriptions given by German speakers in order to determine *applicability conditions* for approximately 20 distinct *situational descriptions* of movement, such as "to go up a street" or "to turn right in front of a landmark". To make clear what we mean with applicability conditions, we present how to describe the "turn-right" action in figure 4. We found that speakers normally use one landmark in order to describe a turning movement. Thus, the applicability conditions for the specific situational description "to turn right in front of L_4" can be formalized as follows (see tabular 2):

situational description	applicability conditions
turn-right in-front-of L_4:	`in-front-of`$(S_1, C_1, da_1) \wedge$ `right-of`$(S_4, C_1, da_2) \wedge$ `in-front-of`$(S_4, L_4, da_3) \wedge$ `in-front-of`$(C_1, L_4, da_4) \wedge$ `on`(P, S_1, da_5)

Tabular 2: Situational description

Applicability conditions consist of conjunctions or disjunctions of spatial relations. Associated with each spatial relation is a limit of applicability. If, for instance, the minimum of the spatial relation `in-front-of`(S_1, C_1, da_1) is greater than the degree of applicability da_1 then the spatial relation is not applicable in this situation. The applicability condition is only satisfied if all spatial relations on the right side are satisfied. In figure 4, the applicability conditions for "turn right in front of L_4" are satisfied if we assume that in this example all $da's$ are equal 0.5. Thus, the situation can be described by translating a specific situational description into verbal descriptions. For this, we use an incremental[6] speech generator which is based on the tree adjoining grammar approach (cf. [Finkler & Schauder 92]).

[6] With incrementality is meant, that parts of a sentence are presented although the sentence is not completely determined.

5 Conclusion

What we can talk about depends on what we see in the current environment. Considering this, we have presented an approach of how we can integrate layout information obtained by visual perception and natural language. We have focused on the representation of visually perceived objects, on the computation of spatial relations concerning their basic meanings, functional dependencies, and the influence of perspective view, and the use of layout information in determining descriptions of the current environment. The area for our model is incremental route description.

An open question is whether our set of applicability conditions are sufficient for all kinds of situational descriptions. Furthermore, we will examine how dynamic objects, such as cars, affect incremental route descriptions. Associated with this is the question of how we can integrate temporal constraints into situational descriptions.

6 Acknowledgements

We are grateful to the following people for valuable discussion: Jörg Baus, Gerd Herzog, Anthony Jameson, Joachim Paul, and Jörg Schirra. Specifically we would like to thank Amy Norton for improving the readability of this paper. Responsibility for infelicities remains entirely our own. Authorship of this paper is shared equally.

References

[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.

[Aurnague et al. 90] M. **Aurnague**, M. **Borillo**, and L. **Vieu**. *A Cognitive Approach to the Semantics of Space.* In: COGNITIVA 90, 1990.

[Bierwisch & Lang 89] M. **Bierwisch** and E. **Lang** (eds.). *Dimensional Adjectives: Grammatical Structure and Conceptual Interpretation.* Berlin-Heidelberg-New York: Springer-Verlag, 1989.

[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 94a] K.-P. **Gapp**. *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.

[Gapp 94b] K.-P. **Gapp**. *From Vision to Language: A Cognitive Approach to the Computation of Spatial Relations in 3D Space.* submitted for publication, 1994.

[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.

[Grice 75] H. P. **Grice**. *Logic and Conversation.* In: P. Cole and J. L. Morgan (eds.), Speech Acts, pp. 41–58. London: Academic Press, 1975.

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

[Kochen 74] M. **Kochen**. *Representations and Algorithms for Cognitive Learning.* Artificial Intelligence, 5:199–216, 1974.

[Kosslyn et al. 90] S. **Kosslyn**, R. **Flynn**, J. **Amsterdam**, and G. **Wang**. *Components of High-level Vision: A Cognitive Neuroscience Analysis and Accounts of Neurological Syndromes.* Cognition, 34:203–277, 1990.

[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.

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

[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.* AI Review Journal, 1994. Special Volume (Issues 1,2,3): Integration of Natural Language and Vision Processing, forthcoming.

[Marr 82] D. **Marr**. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: Freemann, 1982.

[Miller & Johnson-Laird 76] G. A. **Miller** and P. N. **Johnson-Laird**. *Language and Perception.* Cambridge, London: Cambridge University Press, 1976.

[Pinker 84] S. **Pinker**. *Visual Cognition: An Introduction.* Cognition, 18:1–63, 1984.

[Schirra 90] J. R. J. **Schirra**. *A Contribution to Reference Semantics of Spatial Prepositions: The Visualization Problem and its Solution in VITRA.* In: C. Zelinsky-Wibbelt (Hrsg.), The Semantics of Prepositions – From Mental Processing to Natural Language Processing, pp. 301–311. Berlin: Mouton de Gruyter, 1990.

[Yamada et al. 88] A. **Yamada**, T. **Nishida**, and S. **Doshita**. *Figuring out Most Plausible Interpretation from Spatial Descriptions.* In: Proc. of the 12 th COLING, pp. 764–769, Budapest, Hungary, 1988.