ANTLIMA - A Listener Model with Mental Images*

Jorg R.J. Schirra, Eva Stopp
SFB 314, VITRA
Fachbereich 14 - Informatik
Universitat des Saarlandes
D-6600 Saarbrucken 11
Federal Republic of Germany

Abstract

Starting from the thesis that the audience expects the speaker to mean the most typical case of the described class of events or situations with respect to the communicated context, we explain a mechanism for representing and using typicality distributions of static spatial relations which is related to Herskovits' analytical framework. Extended to restrictions of speed and temporal duration, this mechanism also allows us to construct dynamic mental images corresponding to the referents of objective sports reports.

1 Introduction

The project VITRA which started in 1985 as part of the German special collaboration programme SFB 314, *AI & Knowledge- Based Systems,* examines the relations between speaking and seeing: a completely operational form of reference semantics for what is visually perceived *is* to be developed. CITYTOUR and SOCCER are two systems constructed in VITRA which broadly speaking - demonstrate the transformations of visual perceptions into language. Here, we will concentrate on SOCCER and its listener model ANTLIMA (for CITYTOUR cf. [Schirra *el* al., 1987]).

2 SOCCER and ANTLIMA

SOCCER simultaneously analyses and describes short soccer scenes in German similar to a live radio report, i.e., simultaneously and in an objective manner, to an audience which is not able to see the game. As input, SOCCER receives data which is generated by the motion analysis system ACTIONS from the signal of a video camera (cf. [Herzog *et* al., 1989]). This Mobile Object Data - *MOD* - consists of the set of the two-dimensional spatial locations and the velocity vectors of every mobile object perceived in the soccer field from a bird's eye view. At every time quantum, ACTIONS delivers such a set. At present, all mobile objects are perceived as mere ideal points. The MOD implicitly refers to the geometry of the football field and its parts given to the system as Static Background *StaB.*

SOCCER does not know the whole scene at once. Like a radio reporter, it has to consider the events during its occurrence. Therefore, all processing steps have to be done incrementally in a kind of pipelining: a selection of already recognized events is verbalized simultaneously with further event recognition. Beside ACTIONS, the architecture of the system consists, broadly, of the Core System including three components (cf. Fig. 1, [Herzog *et* a/., 1989], and [Andre *et* al., 1989]): the component for Event Recognition produces a set of propositions interpreting the given 'percepts' as instances of spatial and spatio-temporal relations - the former essentially correspond to static spatial predicates like 'being to the left of, the latter to predicates of motion events, e.g., 'doing a double pass'; the component Selection chooses some of those event propositions while planning the continuations of the report; the component

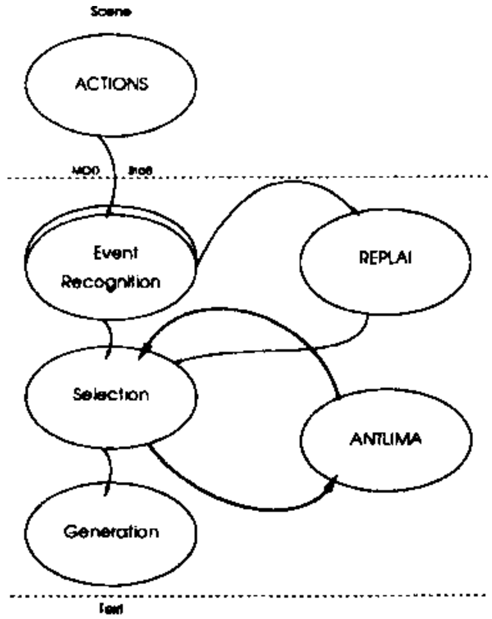

Figure 1: Extended Architecture of SOCCER

*This is an abridged version; the original text is available as VITRA-Report No. 94;

Generation transforms the selected event proposition to an utterance in German coherent to the preceding text, thereby occasionally asking for further information for filling optional deep cases. Whereas the core system recognizes purely spatio-temporal events, SOCCER, has been extended by the component REPLAI which recognizes plans and intentions of the observed agents (cf. [Retz-Schmidt, 1991]).

In this presentation, we are interested essentially in another component of SOCCER,, namely its listener model ANTLIMA. In order to follow (price's Cooperative Principle "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (cf. [Grice, 1974]) , a speaker has to know how his utterance is understood by the listeners in the present context. That means he needs a model of the audience, e.g., to make sure that the listeners are still able to recognize all the relevant information even from elliptic descriptions. Such a model allows him to balance between the divergent demands of economy and completeness - "make your contribution as informative as is required, but also not more informative than is required" - since the speaker can rate how much explicit information actually is required in the given case, and what may be additionally inferred as implicature.

Correspondingly, SOCCER also needs a component that can construct and maintain a model of the listeners' knowledge of the events that have already been described. This listener model enables the system to continue its description in a cooperative way by anticipating the listeners' understanding of the utterance just planned. With these anticipations, the understandability and plausibility of that utterance in the context already known can be rated and used in an anticipation feedback loop to the component, Selection for improving the coherence (cf. Fig. 1). We assume that the understanding of the audience is explained similarly to the text generation of the speaker by reference semantics: the generation of the report depends essentially on the speaker's perception of the referents. Therefore, the listeners also have to represent mentally the referents of the report, albeit referents which they in fact cannot, perceive. As a German linguist wrote in 1969, "the radio reporter has solved his task only if he describes the reality of a sports event so vividly and obviously to the listener that the listener believes he sees that reality." (cf. [Dankert, 1969, p. 94]). The reporter should induce - so to speak a cinema in the heads of his audience. I.e., if the listeners want to have a 'deep' understand ing of an utterance, they should be able following the approach of reference semantics - to create (visual) mental images corresponding to the speaker's percepts (cf. [Schirra, 1990a]).

Correspondingly, the listener model of SOCCER, named ANTLIMA - ANTicipation of the Listeners' 1M-Agery -, must be able to construct visual pseudo-percepts - albeit in the limited sense of SOCCER as MOD and StaB. The abstractly described situation must be concretized again. It is our thesis that those pseudo-percepts generated by ANTLIMA correspond to the listeners' visual mental images, and that they can be used to explain the coherence of the text produced. In the context of listener modeling, the concept 'mental image' in fact participates in two alternative but related frameworks (cf. [Schirra, 1992]): conceived as the mental representation of the understanding of the previous text,, it, allows for explaining the success or failure of acts of reference: an NP be it anaphoric or elliptical can only be used if it uniquely identifies its referent in the image. For example, the NP 'the goal' in a sentence like 'Miller runs toward the goal' is ambiguous if the context does not provide relevant information about the concrete position and direction of movement of Miller which selects one of the two goals of the soccer field. Therefore, the modeling of the listeners' mental images plays an important role for generating adequate noun phrases, especially in the form of elliptical definite descriptions and anaphora ([Schirra, 1991]).

Conceived as the speaker's anticipation of the listeners' understanding of a sentence still to be uttered, the concept 'mental image' makes possible to explain the se lection of the predication: whereas the identification of referents makes use of information assumed to be known already by the audience, the selection of the predication to be communicated is concerned with what the audience presumably does not know yet. The corresponding situation has first to be established in a mental image, thereby elaborating implications specific to the given context. It is this framework and its central question how to construct a mental image corresponding to some given predications which we shall examine in this paper.

ANTLIMA starts with the event proposition chosen by the selection component of SOCCER, as the core of the planned utterance: in order to be made concrete, this proposition has to be analysed down to the primitive propositions: the definition of the considered event concept, i.e., its subevents with their temporal relations, must be expanded and adapted to the situational constraints in order to achieve spatio-temporal coherence between the required spatial restrictions of the involved objects during each phase of the event and the context ([Schirra, 1990b]). Further constraints given by optional deep case fillers also have to be associated to the appropriate phase of the event (cf. [Marburger and Wahlster, 1983] and [Sablayrolles, 1991]). The construction of a corresponding dynamic mental image is performed in two steps: first, the duration of the elementary subevents are fixed; this results in the propositional elementary structure, the temporally ordered sequence of sets of elementary spatio-temporal relations. Each set restricts the scene for exactly one instant (time quantum), describing, so to speak, a 'snapshot'. This data is transformed in a second step to a sequence of MOD relative to the given StaB by fixing particular locations and velocities for all objects considered for each time quantum successively.

In order to rate the understandability of the considered continuation of the report, the generated 'mental image' should not be compared coordinate by coordinate with the original percept of SOCCER: we cannot expect that the listeners will reconstruct the very

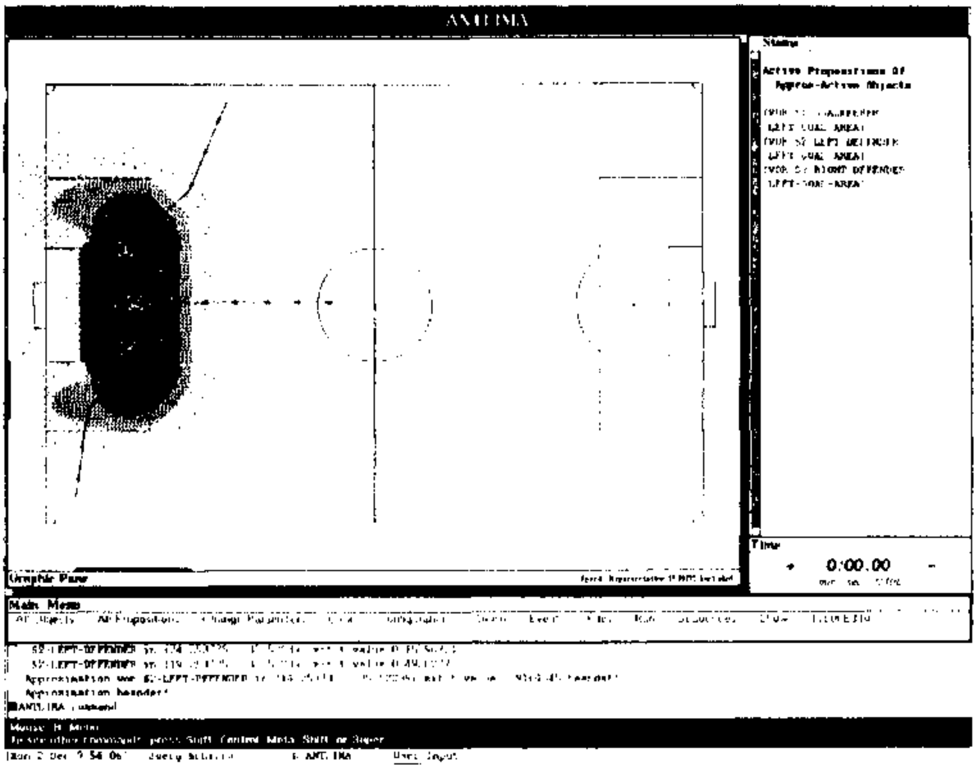

Figure 2: Typicality distribution of the spatial relation 'player in front of the goal area'

same picture from that selected set of propositions actually communicated, since we usually lose information by transforming an image to a description (i.e., propositions) — information which has to be recovered while reconstructing the image. But which deviations are to be tolerated and which are not? If, for example, a player is standing in the middle of the right field — nobody near him — 50 pixels' difference normally will not matter. Though, if he stands at the edge of the field, or near some other player, or very close to the ball, even 10 pixels' deviation of his location might change the whole interpretation of the scene. Since the propositions resulting from the processing of the component Event Recognition ignore by definition details of the percept irrelevant for soccer games, ANTLIMA firstly has to re-analyse its mental images, i.e., describe them propositionally again. The result is a set of propositions that — from the speaker's point of view — represents explicitly the listeners' comprehension of the meaning of the utterance properly intended, including all implicatures, etc. Comparisons of these propositions with those originally recognized in the percept by the core system, those chosen to be communicated, and those propositionally inferred by ANTLIMA's conceptual analysis lead to several sets of differing propositions: correct or wrong understanding, valid or invalid implicatures, and fitting or erroneous expectations of continuations of the scene all can be distinguished by means of these sets. Since this theme does not affect the discussion in this report, we do not deal further with it here (cf. [Schirra, 1991]).

Finally, plausibility ratings, correctly predicted continuations of the report, and — if necessary — detected misunderstandings are given back to the core system (component Selection) where — while closing the anticipation feedback loop — they may be used to increase the pragmatic quality of the report. Additionally, the creation of the mental image and its re-analysis makes it possible to construct a *visual focus* of the audience: the re-analysis brings into focus objects not explicitly mentioned. In consequence, the Generation component can refer to those objects by means of underspecified noun phrases like elliptical definite descriptions without loss of clarity.

3 The Construction of Mental Images

The central assumption for the construction of mental images in ANTLIMA is the following: the audience expects that the speaker always intends to describe the most typical instance of an event. For example, if the speaker is talking about player A transferring the ball to player B the listener will imagine a direct — i.e., linear — transfer of the ball from A to B without any deviations. If there are any differences to the typical case the speaker has to express them, or he violates the listener's communicative expectations. Typicality, on the other hand, depends on the current context. In our example, if the speaker and the audience mutually know that a hard wind has an influence on the trajectory of the ball the speaker can assume that his audience will imagine the ball flying in a curve from A to B as the most typical instance in this particular situation. He therefore does not need to mention explicitly this deviation of the linear trajectory.

For constructing an adequate mental image, i.e., a maximally typical representation of the situation described or restricted - by a proposition, we first have to generate the valid distribution of typicality for all possible mental images. In ANTFIMA, such distributions are called *Typicality Potential Fields* (TyPoFs), functions mapping locations to typicality values.[1] They are used to find the optimal positions: all objects to be located (OLs) are moved to a position with maximal typicality by means of a *hill-climbing algorithm (d.* [Yamada e *al* 1988]). Figure 2 illustrates an exemplary TyPoF corresponding to a proposition of the type 'player in front of the goal area'. The TyPoF is represented graphically by means of grey values: the darker the position, the more typical it is.[2] Note that different starting positions result in different end positions: this influence of the context of course is a very desirable effect which here is given without additional effort. Furthermore (although not demonstrated in the figure), if the reference object A of an OL B is also located with respect to a third object, the TyPoF of B is moved with A: B is pulled along with A until all objects have reached their optimal positions.

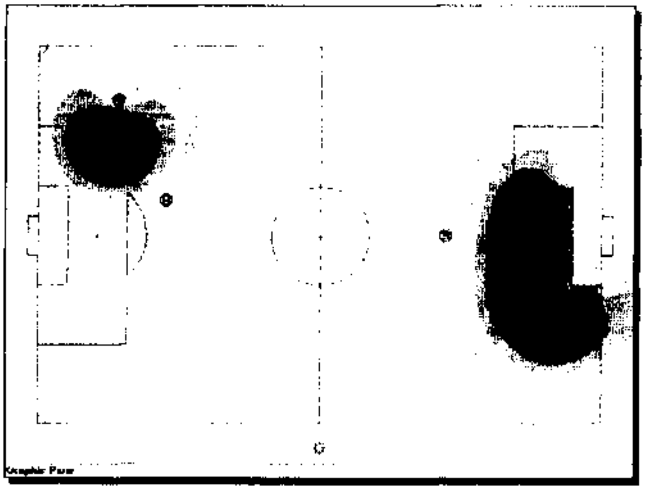

Figure 3 Typicality Distributions of "in front of (extrinsic, use) for Two Different Kinds of Objects

Each TyPoF encodes the typicality distribution corresponding to one kind of proposition, i.e., it depends not only upon the spatial relation used as predicate but also on the arguments of the proposition. Although the typicality distributions of propositions with the same spatial relation, like 'being in front of', but different sets of arguments are more or less similar, the kinds of objects and especially their dimensionality, size, and shape modify the TyPoF of a particular proposition, stretching it, and adaplmg it to the shape of the objects (cf Fig. 3, illustrating the TyPoFs for a player to be in front of player *S2* as seen from the left penalty spot, and for a player to be in front of *the right goal area* as seen from a viewpoint marked with the star). Similar to the system of *'ideal meaning'* and ge*ometrie description functions'* in the analysis of Herskovits (cf. [Herskovits, 1986]), we separate the influence of the objects from the core meaning of the predicate: the assumed particular influence of the arguments, i.e., the objects involved, has to transform the proposed general influence of the spatial relation by adapting it to the shapes, sizes, and possibly also the parts and the functions of the objects involved. To cover the similarity between all instances of a spatial relation, we assume one function which describes the typicality distribution not like TyPoFs with respect to the coordinates of the objects, but with respect to what we call

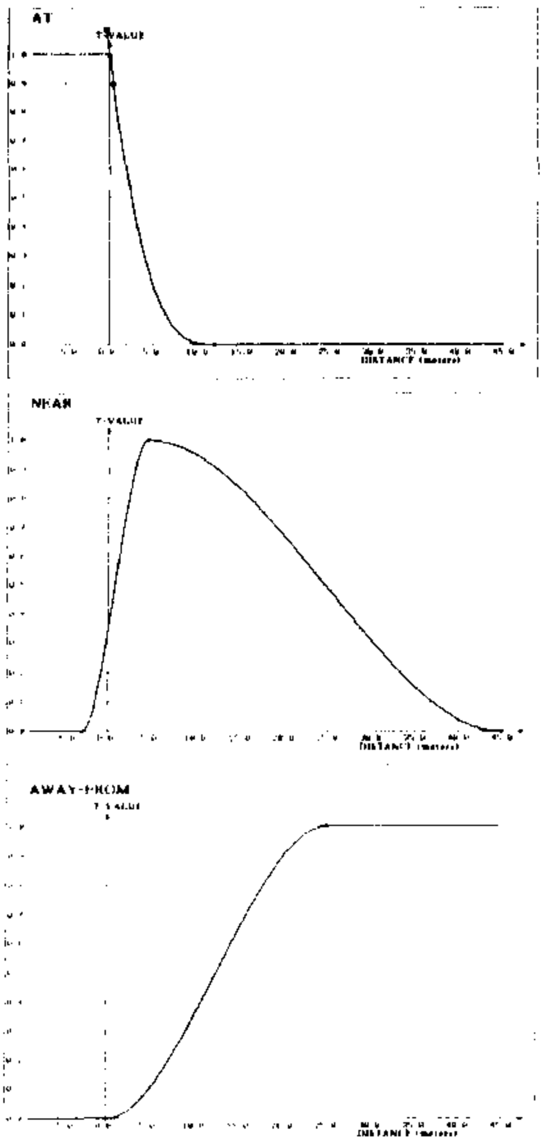

figure 1: (Graphical Representations of the Typicality Schemata of'at "near", and "distant from

'in ANTLIMA, the typicality values are from the rational interval [0.0 ..].()]; the 'T-value' I.O is associated with all locations with maximal typicality;

[2]In Fig. 2, the black parts in fact represent typicality values > 0.9; the propositions used in the exemplary approximation are shown in the Status-Pane on the right side of the figure, in the window on the lower side of the picture, the final two approximation steps for player S2 (black-2) are traced; for a complete list of TyPoFs cf. [Blocher e*t al.*, 1992];

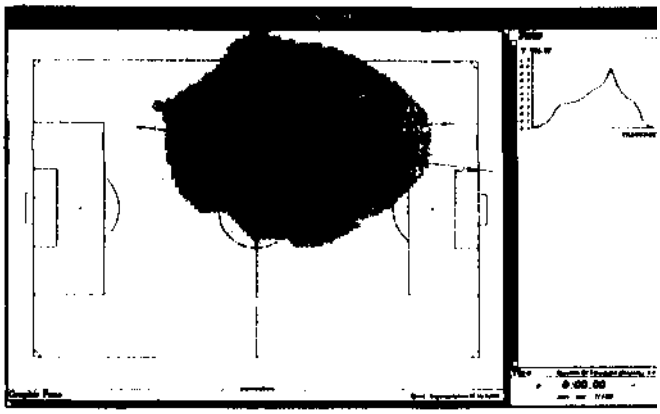

Figure 5: Combination of Four Compatible Restrictions with Sectional View on the Right Side

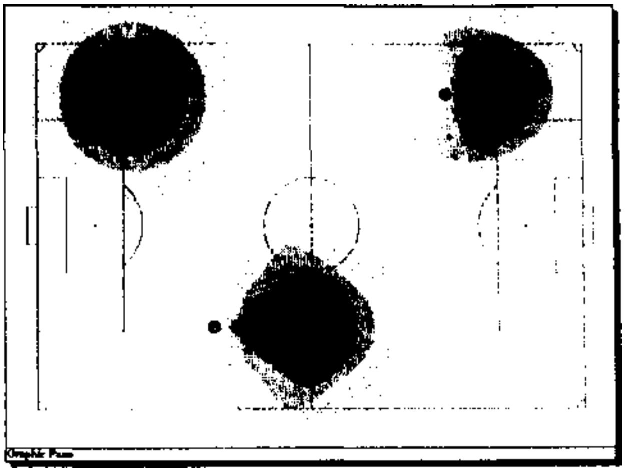

Figure G: TyPoFs of Players' Speeds

essential parameters. The essential parameters, e.g., distance, angles, and scaling factors, are abstractions from the concrete coordinates. From this function the *Typicality Schema* of a spatial relation (cf. Fig. 4) , the TyPoF of any particular instance of that spatial relation is derived. All these TyPoFs instantiated from one Typicality Schema differ only with respect to the functions used to calculate the essential parameters from the objects' coordinates.

The influence of the objects is encoded by ***TyPoF Instantiation Rules*** (for short: ***I-rules***) which in a way ***spread*** the typicality distributions encoded in the Typicality Schema around the objects in the context, thus developing the appropriate TyPoF. Each I-rule specifies, according to the object attributes, a set of functions for calculating the essential parameters. These ***parameter functions*** transform the coordinates of the objects, i.e., the information in the percept, to the essential parameters needed by the Typicality Schemata. Since, for example, ***distance***, the essential parameter used for 'being in', '- at', '- close to', '- near' (cf. Fig. 4), but also involved in 'being left of', '- right of', '- in front of', and '- behind', is defined only between zero-dimensional geometrical objects, we have to reduce or *idealize* higher-dimensional objects to points in order to apply the corresponding typicality schemata. To find the distance between two two-dimensional objects, we first have to look for those two points of the objects' borders which are closest to each other: their distance is the distance between the objects, i.e., they in fact idealize the objects.

If an object is simultaneously restricted by several propositions, the corresponding TyPoFs have to be combined appropriately: algebraic average provides a simple method which also allows us to distinguish compatible and incompatible sets of propositions. The TyPoFs of compatible restrictions interfere in such a way that some locations remain with high typicality in the combined field (≥ 0.8, cf. Fig. 5).[3] Incompatible restrictions result in a combined TyPoF with relatively low maxima reflecting the fact that incompatible restrictions cannot all be fulfilled at once.

To restrict the momentary velocity of an object, the duration of an event, or the temporal relation between events, the concept of TyPoFs was extended: the domain of temporal TyPoFs which are used in the first step of concretizing obviously is only one-dimensional in contrast to the TyPoFs controlling location or velocity. TyPoFs restricting the velocity correspond rather directly to TyPoFs restricting the location of an object: the former bind together two 'incarnations' of the object at consecutive instants. The velocity restriction is interpreted as a restriction of the location at the next moment. Also in these cases, we have to differentiate between the influence of the velocity predicate and the influence of the object: a ball, for example, reaches usually a much higher velocity than a player, a difference illustrated by Fig.s 6 and 7: both figures show the typicality distributions for the three qualitative levels of speed considered in ANTLIMA namely low, medium, and high , the first for a player, the second for the ball (the typicality distributions for the locations after two seconds are shown). Since it is more difficult to change the direction of motion within one time quantum the faster an object moves, the 'angle' of the speed-TyPoFs varies with the speed level.

A sequence of sets of elementary restrictions of location and speed, like the propositional elementary structure analysed from an event proposition, is concretized by reconstructing a corresponding sequence of static situations like a film incrementally moment after moment: the last reconstructed 'snapshot' is used as the context for the concretizing of the next instant by means of combining the velocity restrictions of the already reconstructed time point and the locational restrictions of the instant being considered; the positions of the objects at one moment are used as starting positions for the approximation of their positions one time quantum later (an example is included in the long version). The speed TyPoFs may be used additionally for adequately reduc-

[3] In Fig. 5, the combination of the following local restrictions for B-6 is depicted: (between W-3 Right Goal), (near Center Circle), (at Halfway Line), and additionally ('reachable with maximal speed in 3 sec');

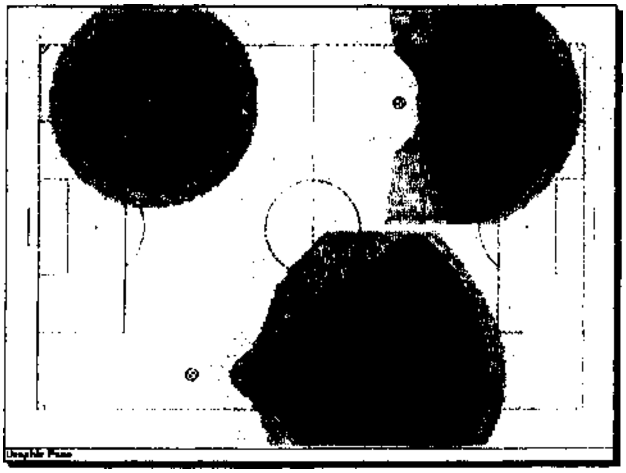

Figure 7: TyPoPs of BalFs Speeds

ing the range of typical locations: those not reachahle for the concerned object within one time quantum inn by maximal speed may I>e optionally faded out' in AN I LIMA.[1]

4 Conclusion

With ANTLIMA, the systematic application of the reference semantics approach for combining vision and natural language systems can be studied: the reconstruction of perceptually absent, referents in the form of mental images is to be used in a listener model for anticipat ing what the audience can comprehend, with the goal of increasing appropriately the coherence of the generated text. We have demonstrated herein the basic principles of such a reconstruction (the system is presently realized on CLOS and CLIM); on this foundation, furl heist tidies, especially concerning the treatment of three-dimensional objects, the conceptualization of movements as deformable one-dimensional objects called 'trajectories[1] and their use (cf. [Hays, 1989]), and the re-analysis of the images with the feedback to text planning and generation, are currently in progress.

References

[Andre et al., 1989] F. Andre'. (.;. Ilerzog, 1. Rist. Natural Language Access to Visual Data: Dealing with Space and Movement. Bericht 03, STB 314. VITRA. Univ. d. Saarlandes, Saarbriicken, 1989.

[Blocher et a/., 1992] A. Blocher, F. Stopp, T.Weis ANTLIMA-1 Fin System zur Cenerierung von Bildvorstellungen ausgehend von Propositionen. Memo 50, SFB 314, VITRA. Univ. d. Saarlandes. Saarbriicken, 1992.

[Dankert, 1909] H. Dankert. Sportsprache and Kommihnikation (Jntersuchungen zur Struktur der EujBballsprache und zum Stil der Sportlnru htirstattung. Vereinigung fiir Volkskunde e.V., 'I'iibingen, 1969.

[Grice, 1974] H.P. (Grice. Logic and Conversation. In P. Cole, J.L. Morgan (eds.), Syntax and Semantics, vol 3, p. 41 58. Academic Press, New York, 1974.

[Hays, 1989] P.M. Hays. Two Yievvs of Motion: on Representing Move Invents in a Language-Vision System. In DMetzing (ed.), GWA1-89,~ p. 312 317, Berlin. 1989. Springer.

[Ilerskovits, 1986] A. Ilerskovits Language and Cognition An Interdisciplinary Study of the Prepositions in English. Cambridge, University Press, 1980.

[Herzog et al., 1989] G. Herzog, C, K. Sung, F. Andre; W. Fnkelmann, H-H Nagel, T. Hist, W. Wahlster, C. Zimniermann. Incremental Natural Language Description of Dynamic Imagery. In W. Brauer, C, Preksa (eds.). W'IssensbasH rtt Systcmi, p. 153 102, Berlin. 19S9. Springer

[Marburger and Wahlster, 1983] II. Marburger, W. Wahlster. Case Hole Filling as a Side Effect of Vi sual Search. In P]roc. EACL-83, p. 188 195. 1983.

[Hetz-Schmidt, 199l] G. Retz-Schmidt. Recognizing Intentions. Interactions, and Causes of Plan Failures. User Modeling and User- Adapted Interaction, 1(1): 173 202, 1991

[Sablayrolles. I99I] P. Sablayroiles. Analyse conceptuelle et representation des verbes de deplacement du francais. lech. Report IRIT/91 23 R, Universite Paul Sabatier, Toulouse. 1991.

[Schirra wt at.. 1987] J.R..I. Schirra, G. Bosch, C K Sung, G. Zimmermann. From Image Sequences to Natural Language: A First Step towards Automatic Perception and Description of Motions. Applied AL I(3):287 305, 1987.

[Schirra. 1990a] J.R.J. Schirra. A Contribution to Reference Semantics of Spatial Prepositions: The Visualizalion Problem and its Solution in VITRA In C /elinsky Wibbelt (ed.), 77K Semantics of /'repositions From Mental Processing TO SIIKJ to Natural languagt processing Moulon de Gruyter, Berlin (to appear in 93) also available as VITRA-Heport No 75. 1990

[Schirra, 1990b] .J.R.J. Schirra. Expansion von Ereiguis Propositionen zur Vsualisierung Die (irundlagen <ler begrilllichen Analyse von AN I LIMA In II. Marburger (ed.), GWAI-90 p. 2 10 250, Berlin. 1990 Springer

[Schirra, 199l] .J.H.J. Schirra Zum Nut/en antizipierter Bildvorstellungen bei der sprachlichen Szenenbeschreibung. VITRA-Memo 49, SFB 314, YITRA, Univ. d. Saarlandes, Saarbriicken, 1991.

[Schirra, 1992] J.R.J Schirra. Connecting Visual and Verbal Space Preliminary Considerations Concern ing the Concept Mental Image'. In M. Borillo, (ed.), Proc. 4th "' STMS Workshop, Toulouse, IRIT, Univ. Paul Sabatier (to appear in 93). also available as VITRA-Report No. 90. 1992.

[Yamada et at., 1988] A. Yamada, T. Nishida, S Doshita. Figuring out Most Plausible Interpretations from Spatial Descriptions. In Proc. 12 th COLINC Budapest, p. 704 769, 1988.

[4]'Io avoid distortions of the typicality distributions, tins option was here only applied for Fig. 5 (cf. footnote 3);