Generating a Coherent Text Describing a Traffic Scene

Hans-Joachim Novak[1]

Fachbereich Informatik, Universität Hamburg
D-2000 Hamburg 13, West-Germany

Abstract

If a system that embodies a reference semantic for motion verbs and prepositions is to generate a coherent text describing the recognized motions it needs a decision procedure to select the events. In NAOS event selection is done by use of a specialization hierarchy of motion verbs. The strategy of **anticipated visualization** is used for the selection of optional deep cases. The system exhibits low-level strategies which are based on verbinherent properties that allow the generation of a coherent descriptive text.

1 Introduction

This contribution focuses on the verbalization component of the NAOS system (the acronym stands for **NA**tural language description of **O**bject movements in a traffic **S**cene). NAOS is designed to explore the border area between computer vision and natural language processing, especially the realm of recognizing and verbalizing motion concepts in image sequences.

NAOS goes all the way from a representation of a real-world traffic scene to a natural language text describing the scene.

The representation of the scene basically consists of its geometry (therefore called geometric scene description (GSD)). To give an impression of the representation a GSD contains for each frame of the image sequence:

- instance of time
- visible objects
- viewpoint
- illumination
- 3D shape
- surface characteristics (color)
- class
- identity
- 3D position and orientation in each frame

(for a detailed description of the GSD see [16]).

For event recognition we use event models ([18], [19]) which define a reference semantic for motion verbs. In the current implementation of the NAOS system about 35 motion verbs and the prepositions **beside**, **by**, **in-front-of**, **near**, and **on** may be recognized by matching the event models against the representation of the scene.

In this paper we are neither concerned with the representation of the underlying scene data nor with the question of event recognition as these issues have been published elsewhere (see [16] [17] [20]). Instead, we focus on the generation of a coherent text describing the image sequence.

In the next section we briefly describe the representation of the recognized events which form the initial data for the verbalization component. Then the overall strategy for composing a coherent description is discussed. The following section introduces a partial solution to the selection problem which is based on the strategy of anticipated visualization. Fourth, we show how some linguistic choices like **passive**, **restrictive relative clauses**, and **negation** are natural consequences of the task of generating unambiguous referring expressions. In the last section we relate our research with current work on language generation.

2 Initial Data

Verbalization starts when event recognition has been achieved. Besides complex events like **overtake** and **turn off**, other predicates like **in-front-of**, **besides**, **move**, etc. are also instantiated. Below is a section of the database after event recognition has taken place (the original entries are in German).

1: (MOVE PERSON1 0 40)
2: (WALK PERSON1 0 40)
3: (RECEDE PERSON1 FBI 20 40)
4: (OVERTAKE BMW1 VW1 (10 12) (15 32))
5: (MOVE BMW1 10 40)
6: (IN-FRONT-OF VW1 TRAFFIC-LIGHT1 27 32)

The above entries are instantiations of event models containing symbolic identifiers for scene objects (e.g. BMW1). The last two elements of an instantiation denote the start and end time of the event.

We use the following notations to denote the event time:

1. $(....Tb\ Te)$
2. $(....(Tb_{min}\ Tb_{max})\ (Te_{min}\ Te_{max}))$
3. $(....(Tb_{min}\ Tb_{max})\ Te)$
4. $(....Tb\ (Te_{min}\ Te_{max}))$

Tb, Te denote start and end time of an event. The first notation is used for **durative** events (e.g. **move**). A **durative** event is also valid for each subinterval of $(Tb\ Te)$.

The second notation is used for **non-durative** events (e.g. **overtake**). Start and end time of such an event are both restricted by lower and upper bounds. Note, that **non-durative** events are not valid for each subinterval of the event boundaries.

The third notation is used for **resultative** events (e.g. **stop**). The start time of a **resultative** event lies within an interval whereas the end time is a time-point.

Finally, the last notation is used for **inchoative** events (e.g. **start moving**, corresponding to the German verb **losfahren**). **Inchoative** events have a well defined start time whereas the end time lies within an interval.

For the task of generating a coherent description of a traffic scene NAOS first instantiates all event models and predicates which may be instantiated using the scene data. This leads to the well known selection problem of natural language generation. For one object there may be many instantiations with different time intervals, hence the task of the verbalization component to choose what to say. In the next section we discuss the theoretical background on which our verbalization component is based.

3 Theoretical Background

In general, language is not generated per se but is always intended for a hearer. Furthermore, language is used to fulfil certain

[1] I thank B. Neumann who contributed several ideas to this article.

goals of the speaker which may sometimes simply be to inform the hearer about certain facts.

In NAOS the generation of a description of the underlying image sequence aims at diminishing the discrepancy between the system's knowledge of the scene and the hearer's knowledge (the same motivation is used in Davey's program [6]). Concernig the hearer we make the following assumptions:

1. S/he knows the static background of the scene, i.e. the streets, houses, traffic lights, etc.

2. S/he did not utter specific interests except: *Describe the scene!*

A description may be the result of such diverse speech acts as **INFORM, PROMISE, PERSUADE,** or **CONVINCE**. NAOS only generates the speech act **INFORM**.

To inform a hearer about something means to tell her/him something s/he has not known before, something that is true and new. In NAOS the definition of true utterances builds on the situational semantics of Barwise and Perry [3]. They understand the meaning of an utterance as a relation between the utterance and the described situation. The interpretation of an utterance by a hearer usually consists of a set of possible situations with a meaning relation to the utterance. We now define an utterance to be true if the set of possible situations contains the actually occurred situation.

The requirement to generate true utterances has two consequences for our verbalization component. First, the verbalization process must take the hearer's meaning relations into account. This coincides with the communication rule to tune one's utterances to the hearer's comprehension ability. Second, assuming that the speaker has the same meaning relations as the hearer, the speaker can anticipate the hearer's interpretation of an utterance, i.e. the possible situations implied solely by the utterance can be generated without knowledge of the actual situation. In the case of scene descriptions these situations are equivalent to the hearer's visualization of an unknown scene.

An utterance must be new to the hearer in order to inform him. In the context of situational semantics we define an utterance to be new if its interpretation restricts the set of possible situations implied by previous utterances. Thus new information additionally specifies described situations.

The task of a verbalization component is to choose utterances such that they inform in the above sense. Therefore it is necessary to anticipate the hearer's understanding for judging whether a planned utterance carries new information.

The general principle for hearer simulation is depicted in figure 1.

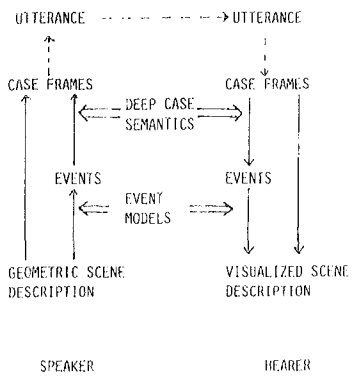

Figure 1: Hearersimulation

On the side of the speaker the event recognition process leads by use of event models to instantiated event models (called events in the figure). A first selection process chooses among the instantiations those which are to be verbalized. As event models are associated with verbs the appropriate case frame of the verb is available. A second selection process now chooses among the optional deep cases of the verb. This is where the deep case semantics comes into play. If, for instance, it is decided that a locative expression should be generated it is necessary to know how the location of an object may be expressed in natural language as in the geometric scene description the location of an object is given by its x, y, and z coordinates. The deep case semantics also contains information about the prepositions which may be used for expressing a specific deep case.

Assuming that the hearer has the same meaning relations as the speaker he basically can use the speaker's processes in reverse order and reconstruct the underlying case frame from the utterance and thus build a visualized scene description.

Note, however, that we agree with Olson [21] that the verbalization of a visual event always leads to a loss of information. In our case, for instance, we cannot assume that the hearer knows the x, y, and z coordinates of an object when he hears the phrase *in front of the department of computer science*. Such a phrase generates a set of coordinates defining the region which corresponds to the preposition **in-front-of**. The actual location of the object which gave rise to the generation of the phrase lies somewhere within that region. Presently, hearer modeling stops at the level of case frames and the visualized scene is anticipated (see section 4.2).

As shown in figure 1 the case frame of a verb plays a central role in our verbalization component. We adopt the view of Fillmore expressed in his scenes-and-frames semantic [7] that case frames relate scenes to natural language expressions.

4 The Selection Problem

Usually this problem is divided into the subtasks of deciding **what** to say and **how** to say it. As mentioned above NAOS uses two selection processes. First, it selects among the instantiated events and second, it selects among the optional deep cases of the verb associated with the chosen event. The first selection process corresponds to deciding what to say and the second one determines largely how to say it as will be shown later.

The selection processes are based on the representation of the case semantics of an event model and on a specialization hierarchy of the verbs. Below is the representation of the case semantics for the event model **überholen (overtake)**.

```
Agent-Restr.:   VEHICLE
  Deep-cases:   (VERB UEBERHOL)
                (UEBERHOLEN *OBJ1 *OBJ2 *T1 *T2)
  Obligatory:   (AGENT AGT-EXP)
                (REF AGT-EXP *OBJ1)

                (TENSE TNS-EXP)
                (TIME-REF TNS-EXP *T1 *T2)

                (OBJECTIVE OBJ-EXP)
                (REF OBJ-EXP *OBJ2)

    Optional:   (LOCATIVE LOC-EXP)
                (LOC-REF LOC-EXP *OBJ1 *T1 *T2)
Combinations:   NIL
   Loc-preps:   (AN AUF BEI HINTER IN NEBEN
                 UEBER UNTER VOR ZWISCHEN)
```

The first slot specifies the agent restriction. The deep-cases slot

contains first the verb stem of **überholen** as needed by the generation component and second the formal notation for an instantiation. The obligatory cases must be generated but may be omitted in the surface string in case of elliptic utterances whereas optional deep cases need not be generated at all. In the combinations slot it is represented which deep cases may be generated together (e.g. for the verb **fahren (drive)** it is not allowed to generate a single SOURCE but instead SOURCE and GOAL must be generated). The Loc-preps slot specifies the prepositions which may be used with the verb **überholen** to generate locative expressions.

The case descriptions in the obligatory and optional slots consist of two parts: a declaration of an identifier for the case expression on the language side, and a predicate (in general a list of predicates) relating the case expression to the scene data. The most important predicates are REF, TIME-REF, and LOC-REF.

REF generates referring phrases for internal object descriptors like BMW1. TIME-REF generates the tense of the verb. As descriptions are usually given in present tense, presently TIME-REF only generates this tense. LOC-REF relates the abstract location of the object as given by its coordinates to a natural language expression for a reference object. Note, that REF has to be used to generate a referring phrase for the reference object. Consider the sixth entry of the database in section 2. The instantiation only contains internal identifiers for objects, like traffic-light1, for which referring phrases have to be generated (see section 4 for further details on REF).

In NAOS we use a specialization hierarchy for motion verbs. This hierarchy is pragmatically motivated and is rooted in situational semantics. It is no hierarchy of motion concepts as the one proposed in [23]. It connects general verbs with more special ones. A situation which may be described using a special verb implies the application of all more general verbs. Take for instance the verb **überholen (overtake)**. It implies the use of the more general verbs **vorueberfahren, vorbeifahren (drive past)**, **passieren (pass)**, **naehern-r (approach)**, **entfernen-r (recede)**, **fahren (drive, move)**, and **bewegen-r (move)**.

It should be intuitively plausible that such a hierarchy is also used for event recognition. If, for instance, no **naehern-r (approach)** can be instantiated the more special events need not be tested.

4.1 Event Selection

In NAOS the overall strategy for generating a descriptive text is as follows:

- Group all moving objects according to their classmembership;
- For each object in each group describe the motions of the object for the time interval during which it was visible in the scene.

Event selection for an object is done according to the following algorithm:

1. Collect all events in the interval where the object was visible and where the object was the agent;
2. determine for each timepoint during the object's visibility the most special event of the above collected ones;
3. if two events have the same specificity then either take the one which started earlier and has the same or longer duration as the other one or take the one with longer duration;
4. put the selected events on the verbalization list of the object in temporally consecutive order.

Consider the following example. All events which were found for PERSON1 are:

```
(RECEDE PERSON1 FBI 20 40)(ENTFERNEN-R PERSON1 FBI 20 40)
(WALK PERSON1 0 40)       (GEHEN PERSON1 0 40)
(MOVE PERSON1 0 40)       (BEWEGEN-R PERSON1 0 40)
```

The above algorithm leads by use of the specialization hierarchy to the following verbalization list for PERSON1:

```
(((WALK PERSON1 0 40) (0 20))
((RECEDE PERSON1 FBI 20 40) (20 40)))
```

(The last entry in parenthesis of each selected event denotes the interval in which the event was the most special one.)

4.2 Selection of Optional Deep Cases

This selection process is our first implementation of the strategy of anticipated visualization. The underlying question is: *Which optional deep cases should be selected to restrict the hearer's possibilities of placing the trajectory of an object in his internal model of the static background of the scene?*

In NAOS the selection algorithm answering the above question is rather straightforward. It is based on the manner of action of the verb, the verbtype, and the hearer's knowledge. The algorithm is graphically represented in figure 2.

EVENTTYPE	VERBTYPE	DEEP CASES
NON-DURATIVE, INCHOATIVE	DIR	LOCATIVE?
	LOC	DIRECTION?, LOCATIVE?
	REO	DIRECTION?
RESULTATIVE	DIR,LOC	LOCATIVE?
	REO	NIL
DURATIVE	REO	NIL
$T_{beg} = SB \wedge T_{end} = SE$	DIR, STAT	LOCATIVE?
	LOC	DIRECTION?, LOCATIVE?
$T_{beg} = SB \wedge T_{end} \neq SE$	STAT	LOCATIVE?
	DIR, REO	NIL
	LOC	DIRECTION?
$T_{beg} \neq SB \wedge T_{end} = SE$	STAT	LOCATIVE?
	DIR	LOCATIVE?
	LOC	SOURCE?, DIRECTION?
	REO	NIL
$T_{beg} \neq SB \wedge T_{end} \neq SE$	REO	NIL
	DIR, STAT	LOCATIVE?
	LOC	SOURCE?, GOAL?

Figure 2: Selection of Deep Cases

The abbreviations denote: T_{beg}, T_{end}: start, end time of the event; SB, SE: scene begin and scene end; DIR, LOC, STAT, REO: directional (turn off, return), locomotion (walk, overtake), and static (stand, wait) verbs, finally verbs whose recognition implies reference objects (reach s. th., arrive at).

The figure has to be read as follows. If an inchoative event like **losfahren (start moving)** has to be verbalized which has the verbtype **locomotion**, then choose direction? and locative? as deep cases. The question mark generally means, look into the partnermodel to see whether this deep case has already been generated for another event. If so, determine by use of the object's actual location (represented in the scene representation) whether it is still valid. If this is the case don't generate a natural language expression for this deep case, otherwise do.

Presently the partnermodel contains information about the static background of the scene and about what has been said so far in the same relational notation as was shown for instantiations in section 2. It is being updated when an event is verbalized.

Note, that for durative events the decision is based on whether the start and end time of the event coincide with the beginning or ending of the image sequence. Consider the first case for durative events as given in figure 2. Right from the beginning of the sequence there is a car moving along a street until the sequence ends. In such a case it is not possible to verbalize a source as the object may have started its motion anywhere. To restrict the hearer's visualization,

direction and locative cases are verbalized, leading to a sentence like: *The car moves on Schlüterstreet in direction of Hallerplace.* Verbalizing a direction when the static background is known restricts the trajectory to being on one side of the road. Basically, our direction case is a goal or source case where only two prepositional phrases are allowed, the German phrases **in Richtung** and **aus Richtung (in direction, from direction)**. These phrases do not imply that the motion ends at the goal location as do most prepositional phrases in German which have to be in accusative surface case to denote a goal. The English language is in this respect inherently ambiguous. In the sentence *The car moves behind the truck*, the phrase **behind the truck** may denote a locative or goal deep case. In German these cases are distinguished at the surface. For locative the above sentence translates to *Das Auto fährt hinter* **dem** *LKW*, for the goal case, it translates to *Das Auto fährt hinter* **den** *LKW*.

We have to distinguish different verbtypes as e.g. the meaning of a directional phrase changes with the verbtype. Consider the sentences *The car moves in direction of Hallerplace* versus *The car stands in direction of Hallerplace* (in German both sentences are well formed). The first sentence denotes the direction of the motion whereas the second one denotes the orientation of the car. We thus distinguish between static (STAT) and locomotion (LOC) verbs. The third verbtype, directional (DIR), is used for verbs with a strong directional component like **umkehren (return)**, **abbiegen (turn off)**, etc. As they already imply a certain direction the additional verbalization of a direction using a prepositional phrase does usually not lead to acceptable sentences. The fourth type (REO) is used for verbs like **erreichen (reach s. th.)** having an obligatory locative case.

The main result to note here is that the selection processes are low-level and verboriented. The only higher level goal is to inform the hearer and to convey as much information about an event as possible. In the next section we show by different verbalizations of the same scene how rather complex syntactic structures arise.

5 Generation

The general scheme for the generation process is as follows:

1. Sort the objects according to their classmembership, vehicles first, then persons;
2. in the above partial order sort the objects according to their time of occurrence in the scene, earliest first;
3. do for all elements in each verbalization list of each object
   (a) if the current event has a precedent and its event time is included in the precedent's, begin the sentence with **dabei (in the meantime)**; go to (c);
   (b) if the current event has a precedent and its event time overlaps the precedent's, begin the sentence with **unterdessen** (approx. **in the meantime**); go to (c);
   (c) determine the optional deep cases and build a simple declarative sentence by using all chosen deep cases and applying the deep case semantics.

Two temporally consecutive events are not verbalized using a temporal adverb as in the cases of inclusion and overlapping. This is due to the fact that from the linear order of the sentences the hearer usually infers consecutivity.

The result of the above algorithm is a formal representation of the surface sentence which, roughly, contains the verb's stem, genus verbi, modality, and person, all deep cases in random order, and all stems of the lexical entries which appear in the surface sentence. This representation is taken as input by the system SUTRA (for further details on the formal representation and the SUTRA system see [4]) which then generates a correctly inflected German sentence.

Below is an example of the output of NAOS.

18. ,ausgabe text

DIE SZENE ENTHAELT VIER BEWEGTE OBJEKTE: DREI PKWS UND EINEN FUSSGAENGER.

The scene consists of four moving objects: three vehicles and a pedestrian.

EIN GRUENER VW NAEHERT SICH DEM GROSSEN FUSSGAENGER AUS RICHTUNG HALLERPLATZ. ER FAEHRT AUF DER SCHLUETERSTRASSE.

A green VW approaches the tall pedestrian from the direction of Hallerplace. It drives on Schlueterstreet.

EIN GELBER VW FAEHRT VON DER ALTEN POST VOR DIE AMPEL. WAEHRENDDESSEN ENTFERNT ER SICH VON DEM GRUENEN VW.

A yellow VW drives from the old postoffice to the traffic light. In the meantime it recedes from the green VW.

EIN SCHWARZER BMW FAEHRT IN RICHTUNG HALLERPLATZ. DABEI UEBERHOLT ER DEN GELBEN VW VOR DEM FACHBEREICH INFORMATIK. DER SCHWARZE BMW ENTFERNT SICH VON DEM GRUENEN VW.

A black BMW drives in the direction of Hallerplace. During this time it overtakes the yellow VW in front of the department of computer science. The black BMW recedes from the green VW.

DER GROSSE FUSSGAENGER GEHT IN RICHTUNG DAMMTOR AUF DEM SUEDLICHEN FUSSWEG WESTLICH DER SCHLUETERSTRASSE. WAEHRENDDESSEN ENTFERNT ER SICH VON DEM FACHBEREICH INFORMATIK.

The tall pedestrian walks in the direction of Dammtor on the southern sidewalk west of Schlueterstreet. In the meantime he recedes from the department of computer science.

19. ,logout

The first sentence above is a standard one having the same structure for all different scenes. The remaining four paragraphs are motion descriptions for the four moving objects.

We now discuss step (c) of the above algorithm in more detail as it covers some interesting phenomena.

Consider the third paragraph describing the motions of the yellow VW. The verbalization list for this object is:

```
(((DRIVE VW1 10 20) (10 25))
 ((RECEDE VW1 VW2 25 32) (25 32)))
```

The beginning (SB) and ending of the sequence (SE) lie at points 0 and 40, respectively. According to the selection algorithm (figure 3) a SOURCE should be verbalized for a durative event with the above event time if the verbtype is LOC. The generation algorithm checks whether the chosen optional cases are allowed for the verb, if so, it is further checked whether the combinations are allowed. As a SOURCE may not be generated alone for a **fahren (drive, move)** event, SOURCE and GOAL are generated.

The fourth paragraph shows the outcome of a deep case selection in which the chosen case is not allowed for the verb. The verbalization list for the black BMW contains only **überholen (overtake)** and **entfernen-r (recede)**.

```
(((OVERTAKE BMW1 VW1 (10 12)(12 32) (10 32))
 ((RECEDE BMW1 VW2 20 40) (32 40)))
```

According to event- and verbtype DIRECTION is chosen as the appropriate deep case. As this case may not be used with the verb **overtake** two sentences are generated, one describing the direction

of the motion and the other one describing the specific event. The second sentence begins with a temporal adverb specifying that both motions occur at the same time. In order to generate the two sentences first the classmembership of the agent of the verb which may not take the chosen deep case is determined. Then the specializationhierarchy is used to go up to either **fahren (drive, move)** or **gehen (walk)** as those verbs may take any deep case. Then the sentences are generated.

Consider the following verbalization list:

```
(((OVERTAKE BMW1 VW1 (0 8) (12 18) ( 0 18))
 ((DRIVE BMW1 0 40) (18 40)))
```

Assuming the direction and location of the motion to be the same as before the algorithm presented so far would generate *A black BMW drives in the direction of Hallerplace. During this time it overtakes the yellow VW in front of the department of computer science. The black BMW drives.*

According to the deep case selection algorithm a DIRECTION and LOCATIVE should be generated for the second event above. As both cases have already been generated with the first event and are still valid the sentence *The black BMW drives* is not generated because before generating a sentence it is checked whether the information is already known to the partner.

5.1 Referring Phrases

In this section some aspects of the referring phrase generator are discussed. As can be seen from the example text objects are characterized by their properties, introduced with indefinite noun phrases when they are not single representatives of a class and they may also be pronominalized to add to the coherence of the text. Therefore we use standard techniques as e.g. described in [8], [9].

We want to stress one aspect of our referring phrase generator, namely its capability to generate restrictive relative clauses with motion verbs. As it may be easily the case that a scene contains two objects with similar properties the task arises to distinguish them and generate unequivocal referring expressions.

It is an interesting fact, that we have several options to cope with this problem which each have their consequences.

One option is to adopt McDonald's scheme of generation without precisely knowing what to say next [13]. According to this scheme two similar objects are characterized in the following way in NAOS. When the first one is introduced it is characterized by it's properties e.g. *a yellow VW*. When the second one has to be introduced, REF notices that a yellow VW is already known to the partner and generates the phrase *another yellow VW*. It starts getting interesting in subsequent reference. The objects are then characterized by the events in which they were involved earlier whether as agent or in another role. This leads to referring phrases like *the yellow VW, which receded from the pedestrian* or *the yellow VW, which has been overtaken*. Note, how passive relative clauses arise naturally from the task of generating referring phrases in this paradigm. The same is also true for negation. Consider the case where the first yellow VW, say VW1, has **passed** an object and the second yellow VW, say VW2, has **overtaken** an object and both events are already known to the partner. If REF has to generate again a referring phrase for VW1 it notices that **pass** is a more general verb than **overtake** and may thus also be applied for the overtake event. It therefore generates the phrase *the yellow VW, which has not overtaken the other object* to distinguish it unequivocally from VW2.

Below is an example of this strategy in a text for the same scene as above. The difference to the first scene is that we replaced the green VW by a yellow one.

10. ,ausgabe text

DIE SZENE ENTHAELT VIER BEWEGTE OBJEKTE: DREI PKWS UND EINEN FUSSGAENGER.

The scene consists of four moving objects: three vehicles and a pedestrian.

EIN GELBER VW NAEHERT SICH DEM GROSSEN FUSSGAENGER AUS RICHTUNG HALLERPLATZ. ER FAEHRT AUF DER SCHLUETERSTRASSE.

A yellow VW approaches the tall pedestrian from the direction of Hallerplace. It drives on Schlueterstreet.

EIN ANDERER GELBER VW FAEHRT VON DER ALTEN POST VOR DIE AMPEL. WAEHRENDDESSEN ENTFERNT ER SICH VON DEM GELBEN VW, DER SICH DEM GROSSEN FUSSGAENGER GENAEHERT HAT.

Another yellow VW drives from the old post office to the traffic light. In the meantime it recedes from the yellow VW which approached the tall pedestrian.

EIN SCHWARZER BMW FAEHRT IN RICHTUNG HALLERPLATZ. DABEI UEBERHOLT ER DEN ANDEREN GELBEN VW, DER SICH VON DEM GELBEN VW ENTFERNT HAT, VOR DEM FACHBEREICH INFORMATIK. DER SCHWARZE BMW ENTFERNT SICH VON DEM GELBEN VW, DER NICHT UEBERHOLT WORDEN IST.

A black BMW drives in direction of Hallerplace. During this time it overtakes the other VW which receded from the yellow VW, in front of the department of computer science. The black BMW recedes from the yellow VW which was not overtaken.

DER GROSSE FUSSGAENGER GEHT IN RICHTUNG DAMMTOR AUF DEM SUEDLICHEN FUSSWEG WESTLICH DER SCHLUETERSTRASSE. WAEHRENDDESSEN ENTFERNT ER SICH VON DEM FACHBEREICH INFORMATIK.

The tall pedestrian walks in direction of Dammtor on the southern sidewalk west of Schlueterstreet. In the meantime he recedes from the department of computer science.

11. ,logout

The consequences of this first option are rather complex syntactic structures which are not motivated by higher level stylistic choices.

Let us now look at a second option which has also been implemented. Experience with the above algorithm for different scenes showed, that if more than two similar objects are in a scene the restrictive relative clauses become hardly understandable. We thus determine how many similar objects there are in the scene before we start the generation process. If there are more than two, REF generates names for them and introduces them as e.g. *the first yellow VW, the second yellow VW* and so on and uses these phrases in subsequent references. An example of this strategy would look like the first example text where the different vehicles are named *the first ..., the second ....* The rest of the text would remain the same.

Taking this option implies leaving McDonald's scheme and approaching to a planning paradigm.

It should be noted here that there is a third option which has hardly been investigated, namely to switch from contextual to cotextual reference as in phrases like *the VW I mentioned last*. We need further research before we can use such techniques effectively.

6 Conclusion and Related Research

We have proposed the scheme of anticipated visualization to generate coherent texts describing real-world events (visual data). The selection algorithms are based on low-level, verbinherent properties, and on a pragmatically motivated verb hierarchy. Together with the verbalization component the NAOS system is now fully operational from event recognition to text generation in the do-

main of traffic scenes. As this domain is rich enough to still pose a lot of problems this opens up the opportunity to integrate higher level strategies for e.g. combining sentences, selecting events, generating deictic expressions, etc.

The main difference between NAOS and other systems for language generation is that we approach the verbalization problem from the visual side and thus are led to use basic selection algorithms. Other systems like TALESPIN [15], KDS [12], TEXT [14], KAMP [1], and HAM-ANS [10] start their processing with language whereas NAOS starts with images. In close connection to our research is the work of [2], [24], [23], [22], and [5]. The first four authors deal with questions of motion recognition and with a reference semantic for motion verbs but are not concerned with text generation. They showed that case frames can be used to generate single utterances. Conklin and McDonald use the notion of salience to deal with the selection problem in the task of describing a single image of a natural outdoor scene.

TALESPIN exemplifies that plans and goals of an actor may form the underlying structure of narratives and may thus be motivation for text generation. In KDS a representation of **what to do in case of fire alarm** is transformed into a natural language text. As the initial representation already contains lexical entries and primitive propositions the task is to organize this information anew so that it may be expressed in an English text. Mann and Moore propose rules for combining propositions and re-edit the text continuously to produce the final version. TEXT generates paragraphs as answers to questions about database structure. McKeown has identified discourse strategies for fulfilling three communicative goals: define, compare, and describe. These strategies guide the generation process in deciding what to say next. McKeown uses the question to determine the communicative goal that the text should fulfil. Research of this kind is very important to clarify the relation between the form of a text and its underlying goals.

One of the domains of HAM-ANS is the kind of traffic scene which is also used in NAOS. In this domain HAM-ANS deals with primarily with answering questions about the motions of objects and with overanswering yes/no questions [25]. The dialogue component of HAM-ANS may be connected to NAOS to also allow questions of the user if the generated text was not sufficient for his understanding. An evaluation of the kind of question being asked by a user may help in devising better generation strategies.

KAMP is a system for planning natural language utterances in the domain of task oriented dialogues. The planning algorithm takes the knowledge and beliefs of the hearer into account. This system shows how a priori beliefs of the hearer may also be integrated in NAOS to generate appropriate referring phrases.

It would be interesting to use a planning component for NAOS which would first determine all deep cases necessary to maximally restrict the visualized trajectory of an object's motion sequence and then try to distribute the cases to the different verbs used in the description in order to generate smooth text.

7 Bibliography

[1] Appelt, D.E., **Planning Natural-Language Utterances to Satisfy Multiple Goals.** SRI International, Technical Note 259, Menlo Park, CA., 1982

[2] Badler, N.I., **Temporal Scene Analysis: Conceptual Description of Object Movements.** Report TR-80, Dept. of CS, University of Toronto, 1975

[3] Barwise, J., Perry, J., **Situations and Attitudes.** Bradford Books, MIT Press, 1983

[4] Busemann, S., **Surface Transformations during the Generation of Written German Sentences.** In: Bolc, L. (ed.), Natural Language Generation Systems. Springer, Berlin, 1984

[5] Conklin, E.J., McDonald, D.D., **Salience: The Key to the Selection Problem in Natural Language Generation.** COLING-82, 129-135

[6] Davey, A., **Discourse Production. A Computer Model of Some Aspects of a Speaker.** Edinburgh University Press, 1978

[7] Fillmore, C.J., **Scenes-and-frames Semantics.** In: Zampolli, A. (ed.), Linguistic Structures Processing. North-Holland, Amsterdam, 1977, 55-81

[8] Goldman, N.M., **Conceptual Generation.** In: Schank, R.C. (ed.), Conceptual Information Processing. North-Holland, 1975, 289-371

[9] von Hahn, W., Hoeppner, W., Jameson, A., Wahlster, W., **The Anatomy of the Natural Language Dialogue System HAM-RPM.** In: Bolc, L. (ed.), Natural Language Based Computer Systems. Hanser/McMillan, München, 1980, 119-253

[10] Hoeppner, W., Christaller, T., Marburger, H., Morik, K., Nebel, B., O'Leary, M., Wahlster, W., **Beyond Domain-Independence: Experience with the Development of a German Language Access System to Highly Diverse Background Systems.** IJCAI-83, 588-594

[11] Jameson, A., Wahlster, W., **User Modelling in Anphora Generation: Ellipsis and Definite Description.** ECAI-82, 222-227

[12] Mann, W.C., Moore, J., **Computer Generation of Multiparagraph Text.** AJCL 7(1), 1981, 17-29

[13] McDonald, D.D., **Natural Language Generation as a Computational Problem: an Introduction.** In: Brady, M., Berwick, R.C. (eds.), Computational Models of Discourse. MIT Press, Cambridge, Mass., 1983, 209-265

[14] McKeown, K.R., **Discourse Strategies for Generating Natural-Language Text.** Artificial Intelligence 27, 1985, 1-41

[15] Meehan, J., **TALE-SPIN.** In: Schank, R.C., Riesbeck, C.K. (eds.), Inside Computer Understanding: Five Programs plus Miniatures. LEA, Hillsdale, New Jersey, 1981, 197-258

[16] Neumann, B., **Natural Language Description of Time-Varying Scenes.** In: Waltz, D. (ed.), Advances in Natural Language Processes. Volume 1 (in press); also as FBI-HH-B-105/84, Fachbereich Informatik, Universität Hamburg, 1984

[17] Neumann, B., **On Natural Language Access to Image Sequences: Event Recognition and Verbalization.** Proc. First Conference on Artificial Intelligence Applications (CAIA-84), Denver, Colorado, 1984

[18] Neumann, B., Novak, H.-J., **Natural Language Oriented Event Models for Image Sequence Interpretation: The Issues.** CSRG Techn. Note # 34, University of Toronto, 1983

[19] Neumann, B., Novak, H.-J., **Event Models for Recognition and Natural Language Description of Events in Real-World Image Sequences.** IJCAI-83, 724-726

[20] Novak, H.-J., **A Relational Matching Strategy for Temporal Event Recognition.** In: Laubsch, J. (ed.), GWAI-84. Informatik Fachberichte 103, Springer, 1985, 109-118

[21] Olson D.R., **Language Use for Communicating, Instructing and Thinking.** In: Freedle, R.O., Carroll, J.B. (eds.), Language Comprehension and the Acquisition of Knowledge. Washington, 1972

[22] Okada, N., **Conceptual Taxonomy of Japanese Verbs for Understanding Natural Language and Picture Patterns.** COLING-80, 127-135

[23] Tsotsos, J.K., **A Framework for Visual Motion Understanding.** CSRG TR-114, University of Toronto, 1980

[24] Tsuji, S., Kuroda, S., Morizono, A., **Understanding a Simple Cartoon Film by a Computer Vision System.** IJCAI-77, 609-610

[25] Wahlster, W., Marburger, H., Jameson, A., Busemann, S., **Overanswering Yes-No-Questions: Extended Responses in a NL interface to a Vision System.** IJCAI-83, 643-646