Generating coherent presentations employing textual and visual material

Elisabeth André, Thomas Rist

Angaben zur Veröffentlichung / Publication details:

André, Elisabeth, and Thomas Rist. 1995. "Generating coherent presentations employing textual and visual material." *Artificial Intelligence Review* 9 (2-3): 147–65. https://doi.org/10.1007/bf00849177.

Generating Coherent Presentations Employing Textual and Visual Material

ELISABETH ANDRÉ and THOMAS RIST

German Research Center for Artificial Intelligence (DFKI), Stuhlsatzenhausweg 3, D-66123 Saarbrücken. Email: {andre, rist}@dfki.uni-sb.de

Abstract. The objective of the work described in this paper is the development of an intelligent generation system which is able to combine textual and visual material. As coherent presentations cannot be generated by simply merging verbalization and visualization results into multimedia output, the processes for content determination, medium selection and content realization in different media have to be carefully coordinated. We first show that multimedia presentations and pure text follow similar structuring principles. Based on this insight, we sketch how techniques for planning text and discourse can be generalized to allow the structure and contents of multimedia communications to be planned as well. In particular, we explain how our approach handles the crucial task of process coordination.

1. INTRODUCTION

Multimedia systems which employ several media such as text, graphics, animation and sound for the presentation of information have become widely available during the last decade. A walk through any computer exhibition shows that almost all companies have enriched their product range with multimedia functionality concerning display, storage, processing and creation of multimedia documents. With regard to the presentation of information, this technology offers not only the choice between media, but also the chance to utilize a combination of several media in which the strength of one medium will overcome the weakness of another.

Automated presentation systems as components of user interfaces to next-generation expert systems, control panels and help systems aim at presentations which are tailored to individual users in particular situations. The fact that it is impossible to anticipate the needs and requirements of each potential user in an infinite number of presentation situations leads to the idea of an intelligent system that automatically generates presentations on the fly in a context-sensitive way. The benefits of using an automated presentation system as back-end to an application program are twofold. While the application developer will be released from the burden of presentation design, the user of an application can

expect intelligible presentations satisfying his individual information needs and style preferences.

Recently, there has been increasing interest in the design of automated presentation systems which take advantage of multimedia technology to make presentations more effective (cf. Arens *et al.* 1993b; Badler *et al.* 1991b; Feiner and McKeown 1991; Marks and Reiter 1990; Maybury 1993; Roth *et al.* 1991; Stock and the ALFRESCO Project Team 1993; Wilson *et al.* 1992; Wahlster *et al.* 1993). The main working steps that such intelligent multimedia systems have to accomplish are:

- **Content selection and organization**
  In some cases, the application system more or less determines the contents, and the presentation system only has to eliminate irrelevant details and to organize the rest. For other applications, it is up to the presentation system to extract the information to be communicated from a knowledge base. When determining the contents of a presentation, the system should follow the maxims of Grice (1975) – which are of course not restricted to natural language. On the one hand, the system has to ensure that all relevant information will be encoded. On the other hand, the user should not be unnecessarily informed about facts he already knows. Content organization is not restricted to medium-specific clustering of information. The system also has to ensure structural compatibility between related presentation parts in different media.
- **Coordinated distribution of information on several media**
  Employing different media when presenting information does not automatically contribute to the success of communication – as the great number of more or less badly designed presentations shows. Examples include conventional paper-printed illustrated documents in which pictures may be superficial or even confuse readers (see also the experimental findings of text-picture researchers discussed in Levie (1987), Molitor *et al.* (1989), and computer-based multimedia presentations which, rather than facilitating comprehension, overload a user's perceptual capabilities). An optimal exploitation of different media requires a presentation system to decide carefully when to use one medium in place of another and how to integrate different media in a consistent and coherent manner. This also includes determining an appropriate degree of complementarity and redundancy of information presented in different media. A presentation that contains no redundant information at all tends to be incoherent. If, however, too much information is paraphrased in different media, the user may concentrate on one medium after a short time and probably overlook information.
- **Medium-specific content realization**
  To optimally exploit the available media, a presentation system must manage the medium-specific encoding of information. A straightforward approach is to rely on dedicated generation components, such as a text generator and a graphics generator, which incorporate design expertise and provide mechanisms for the automated selection, creation and combination of textual and graphical elements. However, such components cannot be used as uni-directional backend systems which only produce textual or pictorial output. Since

the results of the different generators should be tailored to each other, each generator has to know how other generators have encoded information. Therefore, each generator has to provide an explicit representation of its encodings.

- **Laying out the generation results**
  Presentation fragments provided by the generators have to be arranged in a multimedia output. A purely geometrical treatment of the layout task would, however, lead to unsatisfactory results. Rather, layout has to be considered as an important carrier of meaning. For example, two pictures that serve to contrast objects should be placed side by side. When using dynamic media, such as animation and speech, layout design also requires the temporal coordination of output units.

In the WIP project, we aimed at the development of a presentation system that automatically accomplishes the tasks described above. The resulting WIP prototype is able to synthesize presentations that combine textual and visual material, all of which has been generated by the system. A brief overview of WIP can be found under the DFKI site description in this volume. For more details on WIP's central components and the system architecture we refer to André *et al.* (1993), Wahlster *et al.* (1993). In this paper, we focus on the basic structuring principles of our approach to multimedia generation. Furthermore, we sketch how a plan-based approach can be used for the realization of a multimedia presentation system. Finally, we show how to coordinate the processes for content selection, organization and medium selection.

2. METHODOLOGICAL BASIS

Since a lot of progress has been made in natural language generation, we were optimistic about generalizing concepts developed for natural language generation in such a way that they become useful in the broader context of multimedia presentations. Although new questions arise, e.g. how to optimally divide the work between the available presentation media, many tasks in multimedia generation closely resemble problems occurring in natural language generation, in particular, the selection and the organization of the contents of the presentation.

2.1. *Multimedia Generation as a Goal-Directed Activity*

Our approach is based on the assumption that not only the generation of text, but also the generation of multimedia presentations can be considered as a sequence of acts that aims to achieve certain goals (cf. André and Rist 1990). We presume that there is at least one act that is central to the goal of the whole document. This act is referred to as the *main act* (MA). Acts supporting the main act are called *subsidiary acts* (SA).[1] Since main and subsidiary acts can, in turn, be composed of main and subsidiary acts, a hierarchical document structure results. While the root of the hierarchy generally corresponds to a complex

communicative act such as describing a process, the leaves are elementary acts, i.e., speech acts (cf. Searle 1980) or pictorial acts (cf. Kjorup 1978).

In Fig. 1, an example of a document fragment[2] is shown with its intentional structure. The goal of this document fragment is to get the user to remove the cover of the water container of an espresso machine. This goal is seen in the words "Remove the cover." After reading the instruction, the user knows that he is required to remove the cover. However, the instruction on its own does not guarantee that he is willing to accomplish the request and also able to do so. These two goals are to be achieved by means of the picture and the verbal utterance "to fill the water container." We can also associate certain goals with the single picture parts. For example, the two arrows are to ensure that the user knows the trajectory of the object to be manipulated.

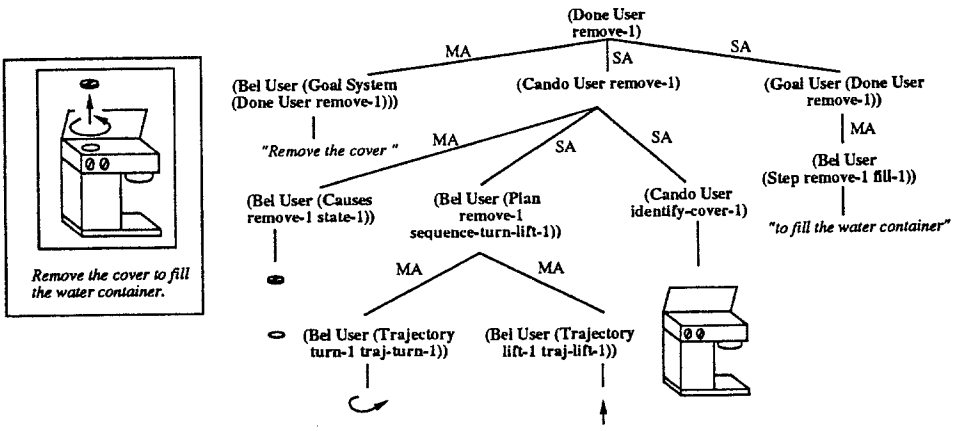

Fig. 1. Intentional structure of a document fragment.

2.2. *An Extended Notion of Coherence for Multimedia Presentations*

A number of text linguists have characterized coherence in terms of coherence relations that hold between the parts of the text (e.g. see Grimes 1975; Hobbs 1978). Perhaps the most elaborate set is presented in Rhetorical Structure Theory (RST, cf. Mann and Thompson 1987), a theory of text coherence. Examples of RST-relations are *Motivation, Elaboration, Enablement, Interpretation* and *Summary*. Psychologists and pedagogues have investigated the role a particular picture plays in relation to accompanying text passages. For example, Levin has found five primary functions (cf. Levin *et al.* 1987): *Decoration, Representation, Organization, Interpretation* and *Transformation*. Hunter and colleagues distinguish between: *Embellish, Reinforce, Elaborate, Summarize* and *Compare* (cf. Hunter *et al.* 1987).

An attempt at a transfer to the relations proposed by Hobbs to pictures and text-picture combinations has been made in Bandyopadhyay (1990). Many psychological and pedagogical studies focus on the various functions of pictures in illustrated documents (cf. Willows and Houghton 1987; Houghton and Willows

1987). Unfortunately, these studies only consider whole pictures, i.e., they do not address the question of how a picture is organized. To get an informative description of the whole document structure, one has to consider relations between picture parts or between picture parts and text passages too. For example, a portion of a picture can serve as background for the rest of the picture or a text passage can elaborate on a particular section of a picture.

Figure 2 shows the rhetorical structure of the document fragment. The document is composed of a request, a motivating part and a part that enables the user to carry out the action. Rhetorical predicates can also be associated with single picture parts. For example, the depiction of the espresso machine serves as a background for the rest of the picture.

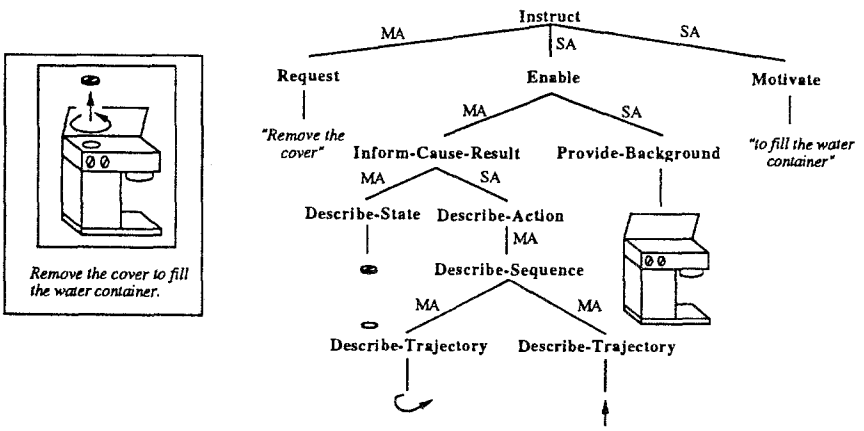

Fig. 2. Rhetorical structure of a document fragment.

2.3. *Cohesive Links between Text and Graphics*

Effective means to establish cohesive links between textual and visual material are referring expressions involving both media. In a multimedia discourse, the following types occur:

Multimedia referring expressions refer to world objects via a combination of at least two media. Each medium conveys some discriminating attributes which in sum allow for a proper identification of the intended object. Examples are natural language expressions that are accompanied by pointing gestures and text-picture combinations where the picture provides information about the appearance of an object while the text restricts the visual search space as in "the switch on the frontside".

Crossmedia referring expressions do not refer to world objects, but to document parts in other presentation media (cf. Wahlster *et al.* 1991). Examples of crossmedia referring expressions are "the upper left corner of the picture" or "Fig. x". In most cases, crossmedia referring expressions are part of a complex multimedia referring expression where they serve to direct the reader's attention to parts of a document that are needed to find the intended referent.

Anaphoric referring expressions refer to world objects in an abbreviated form (cf. Hirst 1981) presuming that they are already explicitly or implicitly introduced in the discourse. The presentation part to which an anaphoric expression refers back is called the antecedent of the referring expression. In a multimedia discourse, we have not only to handle linguistic anaphora with linguistic antecedents, but also linguistic anaphora with pictorial antecedents, and pictorial anaphora with linguistic or pictorial antecedents. For example, the noun phrase "the temperature control" in the first sentence in Fig. 3 refers back to the corresponding switch depiction in the picture (linguistic anaphor with pictorial antecedent). The antecedent of the espresso machine depiction in the picture is the noun phrase "your machine" in the document title (pictorial anaphor with linguistic antecedent). Anaphoric relationships can also be established between picture parts. Examples are the two switch depictions in the inset and the main part of the picture (pictorial anaphor with pictorial antecedent). The example also shows that an anaphor may have more than one direct antecedent. For example, the referring expression "its" in the second sentence is anaphorically connected to the noun phrase "the temperature control" in the first sentence and the switch depiction in the inset.

Examples, such as "the shaded switch," show that the boundary between multimedia referring expressions and anaphora is indistinct. Here, we have to consider whether the user is intended to employ all parts of a presentation for object disambiguation or whether he is supposed to infer anaphoric relations between them.

In André and Rist (1994), we have presented a model of referring which is based on the following assumptions:

1. When a presenter refers to an object, the addressee is intended to activate a mental representation of this object which is already available or which has to be built up (see also Appelt and Kronfeld 1987). Mental representations can be activated not only by textual, but also by graphical or mixed descriptions.
2. Failure and success of referring acts can be explained by the user's ability to recognize certain links between these mental representations and the corresponding object descriptions.

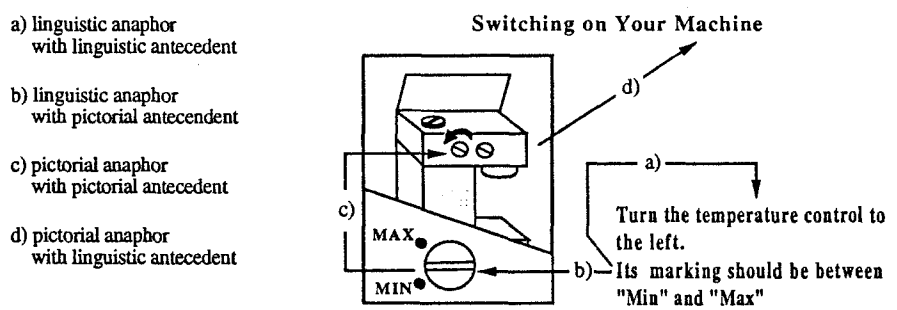

Fig. 3. Different types of anaphora occurring in a sample document.

In addition, the model takes into account that the user's and system's knowledge about the identity of objects doesn't necessarily coincide. For example, the system may believe that the user has different representations for one and the same object without knowing how they are related to each other. In order to describe the links a user has to infer in understanding a referring expression, we have introduced the following predicates:

- The predicate *(Coref repl rep2)* is used to express that two representations *repl* and *rep2* are representations of the same world object.
- The predicate *(Encodes means information context-space)* specifies the semantic relationship between a textual or graphical means and the information the means is to convey in a certain context space.
- The predicate *(EncodesSame p1 p2 context-space)* expresses a cohesive relationship between two presentation parts *p1* and *p2*. It is satisfied if and only if *p1* and *p2* encode the same object.

To illustrate which kinds of coreferential links between referring expressions, world objects and elements of the presentation the user has to infer, let's again have a look at the document fragment shown in Fig. 3. We assume that the user is requested to turn the temperature control. Furthermore, we presume that the user knows of the existence of the on/off switch and the temperature control, has visual access to the two switches, but is not able to tell them apart. Let *r1_u* be a representation for the temperature control the user activates when looking at the picture and *r2_u* be a representation for the temperature control which results from the user's previous knowledge about espresso machines. In the diagrams below, we use the abbreviations *ES, C* and *E* for the *EncodesSame, Coref* and *Encodes* respectively.

To understand the document fragment shown in Fig. 3, the user must be able to infer the anaphoric link between the noun phrase "the temperature control" and the left switch in the main part of the picture. Furthermore, he has to recognize the Encodes-Links between "the temperature control" and *r2_u* and the left switch depiction in the main part of the picture and *r1_u*. From this knowledge, he is expected to infer the Encodes-Link between "the temperature control" and *r1_u* and the Coref-Link between *r1_u* and *r2_u* (cf. Fig. 4a). To understand the second sentence, the user has to infer two anaphoric relationships: the EncodesSame-relationship between the pronoun "its" and the noun phrase "the temperature control" and the EncodesSame-relationship between "its" and the switch depiction in the inset. From these relationships, the Encodes-relationships between "its" and *r1_u* (cf. Fig. 4b) and the switch in the inset and *r1_u* (cf. Fig. 4c) result. Finally, the Encodes-relationships between the two switch depictions and *r1_u* lead to the EncodesSame-relationship between the switch depictions (cf. Fig. 4d).

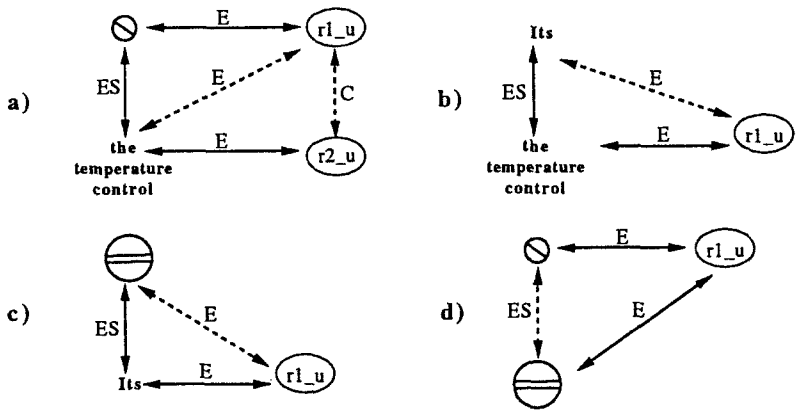

Fig. 4. Referring diagrams for the document fragment shown in Fig. 3.

3. BUILDING A MULTIMEDIA PRESENTATION SYSTEM

3.1. *Planning the Contents and the Structure of Presentations*

As argued in the preceding section, text-picture combinations are similar to text in their structuring principles. In particular, a presentation is characterized by its intentional structure that is reflected by the presenter's intentions and by its rhetorical structure that is reflected by various coherence relations.

Therefore, it seems reasonable to use text planning approaches not only for the organization of the textual parts of a multimedia presentation, but also for structuring the overall presentation. An essential advantage of a uniform structuring approach is that not only relationships within a single medium, but also relationships between parts in different media can be explicitly represented.

To represent presentation knowledge, we have defined presentation strategies which refer to both text and picture production. Each presentation strategy is represented by a header, an effect, a set of applicability conditions and a specification of main and subsidiary acts. Whereas the header of a strategy is a complex communicative act (e.g. to enable an action), its effect refers to an intentional goal (e.g. the user knows a particular object). To represent intentional goals, we use the same notation as in Hovy's RST planner (cf. Hovy 1988). The expression *(Goal P x)* stands for: The presenter *P* has *x* as a goal. *(Bel P x)* should be read as: *P* believes that *x* is satisfied. *(BMB P U x)* is an abbreviation for the infinite conjunction: *(Bel P x) & (Bel P (Bel U x)) & (Bel P (Bel U (Bel P x)))*, etc. The applicability conditions specify when a strategy may be used and constrain the variables to be instantiated. The main and subsidiary acts form the kernel of the strategies. Our representation formalism is similar to that proposed in Moore and Paris (1989). However, we introduce an additional slot for the medium. In doing so, we are able to define medium-independent presentation strategies as well as strategies that apply only for specific medium combinations. Examples of presentation strategies are shown below. Strategy

S1 can be used to introduce an object by showing a picture of it. In this strategy, two kinds of act occur: the elementary act S(surface)-Depict and two complex communicative acts (Label and Provide-Background). The first subsidiary act serves to inform the user about the name of the object, the second is to enable its identification in the picture.

(S1) **Header:** (Introduce System User ?object Graphics)
Effect: (BMB System User (Isa ?object ?concept))
Applicability Conditions:
(Bel System (Isa ?object ?concept))
Main Acts:
(S-Depict System User ?object ?pic-obj ?picture)
Subsidiary Acts:
(Label System User ?object ?medium)
(Provide-Background System User ?object ?pic-obj ?picture Graphics)

To accomplish the second task, strategy S2 may be applied. The main act of this strategy is again an elementary act (S-Depict). Instead of specifying the subsidiary act, only the goal to be achieved is indicated. Whereas the strategy prescribes graphics for the main act, it leaves the medium open for the subsidiary act.

(S2) **Header:** (Provide-Background System User ?x ?px ?picture Graphics)
Effect: (BMB System User (Encodes ?px ?x ?picture))
Applicability Conditions:
(And (Bel System (Encodes ?px ?x ?picture))
(Bel System (Perceptually-Accessible-p U ?x))
(Bel System (Part-of ?x ?z)))
Main Acts:
(S-Depict System User ?z ?pz ?picture)
Subsidiary Acts:
(Achieve System (BMB System User (Encodes ?pz ?z ?picture))
?medium)

To automatically build up presentations, the strategies are considered operators of a planning system. Starting from a communicative goal, the system searches for presentation strategies whose effect subsumes this goal. If such a presentation strategy is found, the expressions in the body are treated as new subgoals. The result of the presentation planning process is a refinement style plan in the form of a directed acyclic graph (DAG). The leaves of this DAG are elementary speech acts or pictorial acts that are forwarded to the medium-specific generators. Note that the medium-specific generators receive not only requests for producing output, but also requests for evaluating their generation results. For example, the subsidiary act of strategy S2 is forwarded to the graphics generator which has to analyze the picture *?picture* to find out whether the object *?z* is identifiable. A further strategy is only applied if the evaluation leads to a negative result. This point underscores the tight connection between image generation and image understanding.

Since there may be several strategies for achieving a certain goal, criteria for ranking the effectiveness, the side-effects and costs of executing presentation strategies are needed. To prioritize presentation strategies, we use selection rules. For example, the selection rule below suggests the use of graphics rather than text when presenting spatial information.

IF (isa ?current-attribute-value SPATIAL-CONCEPT)
THEN (TryBefore *graphics-strategies* *text-strategies*)

Extended studies of relevant psychological literature (e.g., see Willows and Houghton 1987; Houghton and Willows 1987) and our own analyses of various illustrated documents form the basis of our selection rules and presentation strategies. For the generation of instructions, we currently distinguish between 7 information types (concrete, abstract, spatial, covariant, temporal, quantification, negation) with several subtypes and 10 communicative functions (attract-attention, compare, elaborate, enable, elucidate, label, motivate, evidence, background, summarize). To find out which medium or media combination conveys them best, we have analyzed various documents. For example, it is very difficult or even impossible to graphically depict quantifiers (such as *some* or *a few*) whereas graphics are in general the preferred medium for conveying spatial information. Furthermore, we consider criteria such as user characteristics or resource limitations when selecting presentation media. However, the identification of design criteria is an ongoing research area. As more sophisticated models of a user's understanding processes become available, our presentation strategies and selection rules can be refined accordingly.

The plan-based approach is also used for the generation of referring expressions. As in strategy (S3), acts for activating representations may occur in presentation strategies as part of a superordinate speech act.

(S3) **Header:** (Request System User ?action Text)
Effect: (BMB System User (Goal System (Done User ?action)))
Applicability Conditions:
(And (Goal System (Done User ?action))
(Bel System (Complex-Operating-Action ?action))
(Bel System (Agent ?agent ?action))
(Bel System (Object ?object ?action)))
Main Acts:
(S-Request System User (?action-spec (Agent ?agent-spec) (Object ?object-spec)))
Subsidiary Acts:
(Activate System User (Action ?action) ?action-spec Text)
(Activate System User (Agent ?agent) ?agent-spec Text)
(Activate System User (Object ?object) ?object-spec Text)

Strategy (S3) can be used to request the user to perform an action. In this strategy, two kinds of act occur: an elementary speech act (S(urface)-Request) and three activation acts for specifying the action and the semantic case roles associated with the action (Activate). The strategy prescribes text for the subsidiary acts

because the resulting referring expressions (*?action-spec, ?agent-spec* and *?object-spec*) are obligatory case roles of an S-Request speech act which will be conveyed by text. For optional case roles any medium can be taken. To activate representations, strategy (S4) may be applied, which simultaneously enriches the user's knowledge about the identity of objects.

(S4) **Header:** (Activate System User (?case-role ?rep-1) ?spec Text)
Effect: (BMB System User (Coref ?rep-1 ?rep-2))
Applicability Conditions:
(And (BMB System User (Encodes ?pic-obj ?rep-1 ?context))
(Bel System (Coref ?rep-1 ?rep-2))
(Bel System (Bel User (Isa ?rep-2 Thing))))
Main Acts:
(Provide-Unique-Description System User ?rep-2 ?spec Text)
Subsidiary Acts:
(Achieve System (BMB System User (EncodesSame ?spec ?pic-obj ?context)) ?medium)

The strategy only applies if there already exists a picture, if the system knows how the two representations *?rep-1* and *?rep-2* are related to each other and if the system's model of the user's beliefs contains *?rep-2*. If the strategy is applied, the system (a) provides a unique description for *?rep-2* (main act) and (b) ensures that the user recognizes that this description and the corresponding image specify the same object (subsidiary act). For (a), we use a discrimination algorithm similar to the algorithm presented in Reiter and Dale (1992). However, we have investigated additional possibilities for distinguishing objects from their alternatives. We are able to refer not only to features of an object in a scene, but also to features of the graphical model, their interpretation, and to the position of picture objects within the picture (see also Wazinski 1992). A detailed description of our discrimination algorithm can be found in Schneiderlöchner (1994). Task (b) can be accomplished by correlating the visual with the textual focus, by redundantly encoding object attributes, or by explicitly informing the user about a Coref-relationship. Such a Coref-relationship can be established by strategies for the generation of crossmedia referring expressions (as in "The left switch in the figure is the temperature control") or by strategies for annotating objects in a figure. (S5) is an example of such a strategy. It only applies if the system believes that there is a coreferential relationship between *?rep-1* and *?rep-2* and if it is mutually believed that there is a textual element which encodes *?rep-1* and a pictorial element which encodes *?rep-2*.

(S5) **Header:** (Establish-Coreferential-Link System User ?rep-1 ?rep-2 Graphics)
Effect: (BMB System User (Coref ?rep-1 ?rep-2))
Applicability Conditions:
(And (BMB System User (Encodes ?spec-1 ?rep-1 ?text-passage))
(BMB System User (Text-Obj ?spec-1))
(BMB System User (Encodes ?spec-2 ?rep-2 ?picture))
(BMB System User (Pic-Obj ?spec-2)))

Main Acts:
(S-Annotate System User ?spec-1 ?spec-2 ?picture)

3.2. *Process Coordination*

While most people agree on the nature of the decision processes for content selection, content organization and medium selection, architecture models for the processes remain an issue for discussion.

In SAGE (Roth *et al.* 1991), relevant information is selected first and then organized by the text and graphics generators. After that, the generated structures are transformed into text and graphics. A disadvantage of this method is that text and graphics are built up independently of each other.

In COMET (Feiner and McKeown 1991), a tree-like structure that reflects the organization of the presentation to be generated is built up first. This tree is extended by the medium-specific generators in a monotone manner. The system does not allow for revisions caused by medium selection.

Arens and colleagues (Arens *et al.* 1993a) propose a strict separation of planning and medium selection processes. During the planning process, their system fully specifies the discourse structure, which is determined by the communicative goals of the presenter and the content to be communicated. After that, special rules are applied to select an appropriate medium combination. After medium selection, the discourse structure is traversed from bottom to top to transform the discourse structure into a presentation-oriented structure. A problem with this approach is that the presentation structure obviously has no influence on the discourse structure. The selection of a medium is influenced by the discourse structure, but the contents are determined independently of the medium.

In AIMI (Maybury 1993), content selection and content organization are done in parallel. Although AIMI performs medium selection during content selection, dependencies between content selection and medium selection can be handled only to a limited extent. For example, AIMI's operators contain complex communicative acts, such as *Identify (S, H, entity)* which may be realized by either text or graphics. However, since AIMI does not associate medium constraints with such acts, it is not able to express that in a certain context this act should be accomplished by a particular medium.

Our work is distinguished from the work described above in that we use an integrated approach for content selection, content organization and medium selection. The advantage of such an approach is that it facilitates the coordination of the different decision processes. As described in the preceding section, the header of our strategies contains an additional slot for the medium for which constraints can be defined and propagated during the planning process. Depending on whether the main acts of a strategy are to be realized in text, graphics, or both media, the values *Text, Graphics* or *Mixed* are assigned.

The medium remains unspecified until medium decisions are made for the main acts of a strategy. Assume the system decides to compare two objects by describing the different values of a common attribute. At this time, the only restriction is that both descriptions should be realized in the same medium. Once the system has decided on the medium for the attribute value of the first

object, the result of this decision is propagated to the part of the DAG that refers to the description of the second object (cf. Fig. 5).

Since the presentation planner has no direct access to knowledge concerning medium-specific realization, it cannot consider this information when building up a candidate document structure. Consequently, it may happen that the results provided by the generators deviate to a certain extent from the initial document plan. Such deviations are reflected in the DAG by output sharing, structure sharing and structure adding (see also André and Rist 1993). Output sharing occurs when parts of the generated output are reused for different purposes. Structure sharing is similar to output sharing. It occurs when not only parts of the output, but also a more complex part of the DAG is shared. Whereas structure sharing leads to simplifications of the initial document plan, structure adding results in a more complex plan. For example, it occurs if the graphics generator is expected to integrate information in a single picture, but is only able to convey the information by generating several pictures.

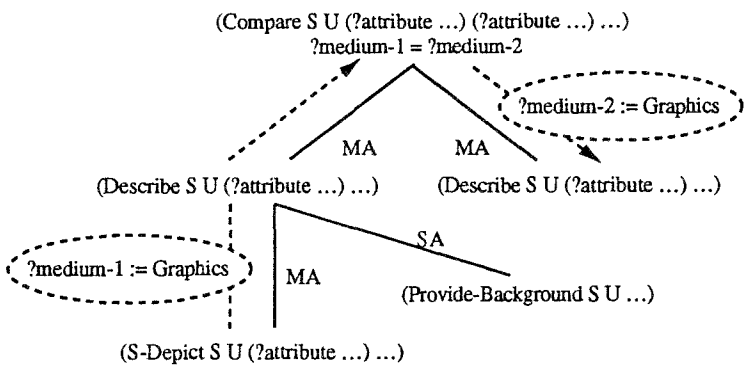

Fig. 5. Propagating medium information.

By means of the following system runs, we illustrate how our system handles the dependencies between content selection, content organization and medium selection. In all system runs, we suppose the system is requested to present a domain plan for setting a modem for reception of data, i.e. the goal *(Instruct System User set-for-reception-1 ?Medium)* has to be accomplished. To show how medium selection influences content selection and content organization and vice versa, we vary the generation parameter 'medium preferences' in each system run.

System run 1: preferred medium graphics

In the first system run, we assume that the user prefers graphics to text. When looking for strategies that match the presentation goal, the system finds two possibilities: Alternative A is to verbally request the user to set a certain code switch to a certain position and to enable him to carry out that action using a medium not yet specified. Alternative B is to describe the action to be carried out and the result to be achieved. Whereas the first alternative prescribes text

for the main act, the second alternative leaves the medium for all acts open. The presentation planner now has to select the strategy which best corresponds to the user's preferences. To prioritize the strategies, we suggest the following heuristic:

Let *PM* be the preferred medium, *S* a strategy, a_i with $1 \leq i \leq n$ main acts of *S*, a_i with $n+1 \leq i \leq n+m$ subsidiary acts of *S*. To each act a_i, we assign a value v_i which depends on the medium the strategy prescribes for a_i. For example, if graphics is preferred to text, as in our case, graphical acts get the value 1 whereas 0 is assigned to textual acts. Acts for which the medium is unknown at definition time get the value 0.5 since the system still has a chance to instantiate them with the preferred medium at runtime. The medium-specific degree of suitability, *MDS(S)*, is defined as follows:

$$MDS(S): = \frac{2 \sum_{i=1}^{n} v_i + \sum_{i=n+1}^{n+m} v_i}{2n + m}$$

$$\text{where } v_i = \begin{cases} 1 & \text{if } S \text{ prescribes } PM \text{ for } a_i \\ 0.5 & \text{if the medium is left open} \\ 0 & \text{if } S \text{ prescribes a medium} \\ & \text{not equal to } PM \text{ for } a_i \end{cases}$$

MDS(S) is a weighted average value of all v_i which expresses how well the strategy *S* satisfies the medium preferences indicated by the user. Starting from the assumption that main acts should have more weight than subsidiary acts, the corresponding values for the main acts in the formula are duplicated.

According to the formula, the MDS for alternative B (0.5) is higher than the MDS for alternative A (0.25). Therefore, B is tried first. Note that the heuristic only considers applicable presentation strategies. In this way, we avoid inadequate media combinations caused by paying too much attention to the user's preferences.

After some expansions, Describe-Orientation and Describe-Trajectory are posted as new subgoals. At this time, the presentation planner instantiates the media. Due to the medium preferences, the presentation planner chooses graphics for both subgoals. The same goes for the Describe-Result goal. Note that medium choices are postponed as long as possible to avoid a situation in which decisions have to be retracted because they are not realizable. Therefore, the medium variable corresponding to the instructing act was not instantiated with graphics immediately after applying strategy B. Figure 6 shows the presentation structure that has been built up so far and the values of the medium slots at the time when the corresponding strategy was applied.

Note that the presentation planner has created three background substructures. When processing these substructures, the graphics generator discovers that it is possible to convey the information requested in a single picture. Consequently, the document structure can be simplified by structure sharing as shown in Fig. 7. This figure also shows the settings of the medium slots after the medium propagation process.

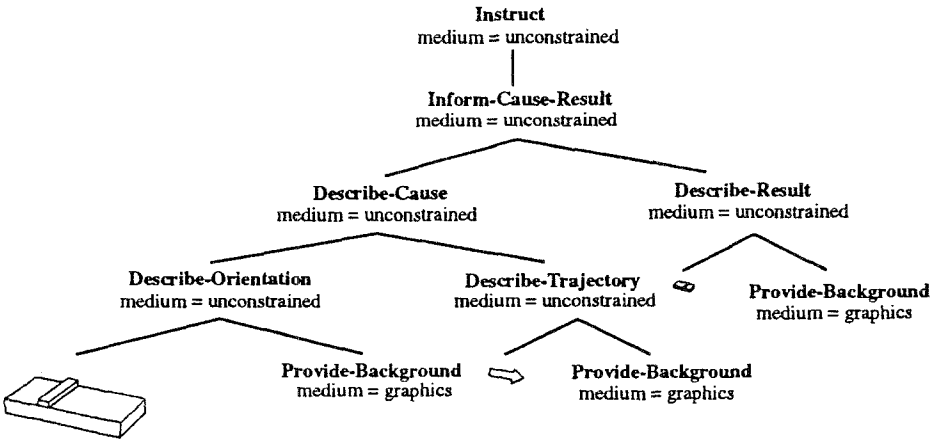

Fig. 6. Preferred medium graphics.

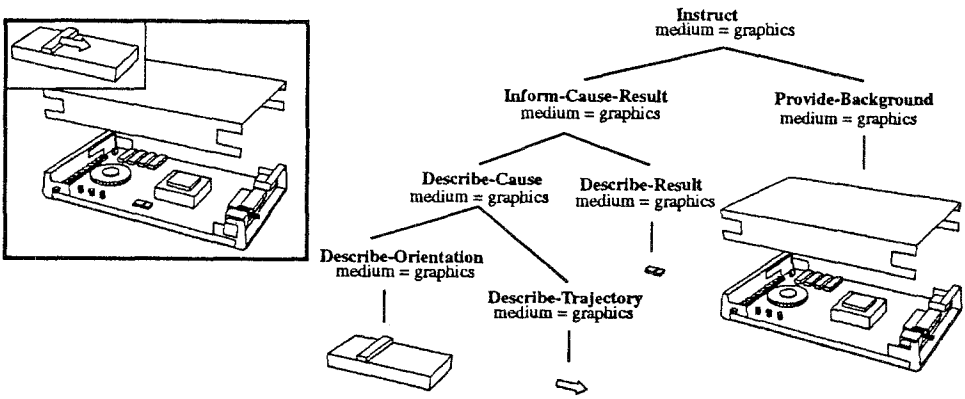

Fig. 7. Presentation structure after factoring out the background subtrees.

System run 2: preferred medium text

In a second system run, we assume that text is the preferred medium. Now, alternative A gets the value 0.83 whereas 0.5 is assigned to alternative B. As a consequence, the system chooses alternative A and verbally requests the user to push the code switch S-4 to the right. As mentioned above, this alternative leaves open which medium should be chosen for the Enable-Act. When looking for strategies to accomplish this act, the system finds two possibilities. The first is to verbally describe the position of all objects involved; the second prescribes graphics for the same task. Since the MDS for the first strategy is higher than the MDS for the second, the presentation planner chooses the first possibility and verbally informs the user where the code switch is located. The presentation structure of this case is shown in Fig. 8.

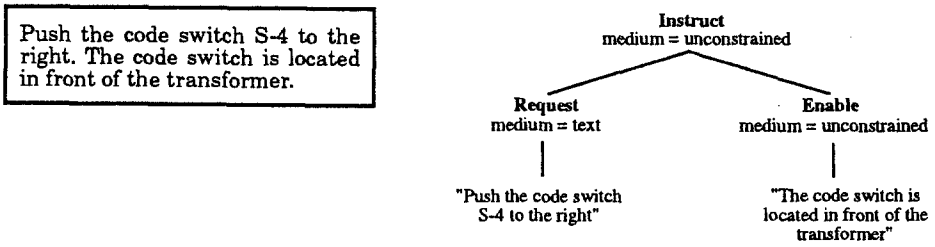

Fig. 8. Preferred medium text.

System run 3: no medium preferences

In a further system run, no specific medium preferences have been indicated. This time, the presentation planner applies possibility A, which is the first strategy that matches the presentation goal. As in the previous system run, the request is conveyed by text. However, since there are no medium preferences, the system follows the selection rule presented in Section 3.1 and uses graphics to describe the spatial location of the switch. As a result, the final presentation consists of a mixture of text and graphics (cf. Fig. 9).

No serial architecture with a total ordering of the components for content selection and content organization would be adequate. Although all system runs serve to accomplish the same instructional goal, the resulting presentation structures differ since they also depend on the medium to be used. Consequently, the system needs information concerning the medium before planning the structure of a presentation. On the other hand, to select a medium the system has to know what to communicate. An essential advantage of our approach is that it allows for more flexibility concerning the timepoint of medium selection. In the extreme case, medium decisions are taken after the complete contents of a presentation are determined. However, it is also possible to select a medium at a very early stage. In this case, the selected medium may influence further content selection

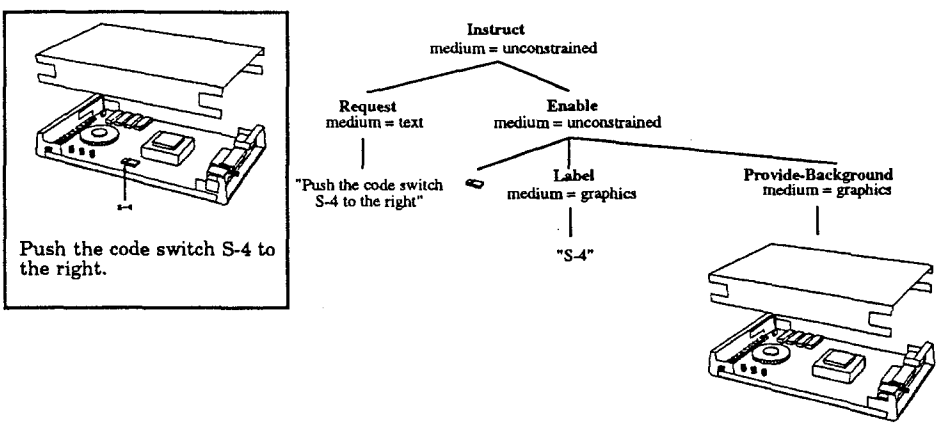

Fig. 9. No medium preferences.

and content organization processes. Despite this flexibility, it may happen that a presentation must be reorganized as in the first example.

4. CONCLUSION

Starting from the insight that multimedia presentations follow similar structuring principles as pure text, we have extended work on text-planning to the broader context of multimedia presentations. We sketched a plan-based approach for the automated synthesis of multimedia presentations. The operators of this planning approach are presentation strategies which refer to both text and picture production. A distinguishing feature of our work is that we use a highly integrated approach for content selection, organization and media selection. A uniform planning formalism facilitates the simultaneous coordination of the central subtasks a multimedia presentation system has to accomplish.

Our approach has been utilized for the implementation of the multimedia presentation system WIP. Depending on the value combination of generation parameters, WIP conveys the same information either with text, graphics, or coherent text-picture combinations.

The design of intelligent multimedia presentation systems is a multidisciplinary endeavor. In particular, this area benefits from research in computer vision and natural language processing. In fact, our work on the design of a multimedia presentation system started with a study of illustrated documents in order to find out how to structure and render a text-picture combination in such a way that an addressee's joint image- and text-understanding processes will eventually lead to the behavior intended by the presenter. As a methodological basis we fell back on well-known concepts from the area of natural language processing like speech acts, rhetorical relations and referring expressions and showed that they take on an extended meaning in the context of multimedia communication.

ACKNOWLEDGEMENTS

The work presented here has been supported by the German Ministry of Research and Technology (BMFT) under grant ITW8901 8. We would like to thank Paul Mc Kevitt and an anonymous reviewer for their helpful comments.

NOTES

[1] This distinction between main and subsidiary acts essentially corresponds to the distinction between *global* and *subsidiary speech acts* in Searle (1980), *main speech acts* and *subordinate speech acts* in van Dijk (1980), *dominierenden Handlungen* and *subsidiären Handlungen* in Brandt *et al.* (1983) and between *nucleus* and *satellites* in RST (Mann and Thompson 1987).

[2] The example is a slightly modified and translated version of instructions for the Philips espresso machine HD 5649.

REFERENCES

André, E. & Rist, T. (1990). Towards a Plan-Based Synthesis of Illustrated Documents. In Proceedings of *The Ninth ECAI*, 25–30. Stockholm. Also as DFKI Research Report RR-90-11.

André, E. & Rist, T. (1993). The Design of Illustrated Documents as a Planning Task. In Maybury, M. (ed.) *Intelligent Multimedia Interfaces*, 94–116. AAAI Press. Also as DFKI Research Report RR-92-45.

André, E. & Rist, T. (1994). Referring to World Objects with Text and Pictures. In Proceedings of *The Fifteenth COLING*, Kyoto, Japan (to appear).

André, E., Finkler, W., Graf, W., Rist, T., Schauder, A. & Wahlster, W. (1993). WIP: The Automatic Synthesis of Multimodal Presentations. In Maybury, M. (ed.) *Intelligent Multimedia Interfaces*, 75–93. AAAI Press. Also as DFKI Research Report RR-92-46.

Appelt, D. & Kronfeld, A. (1987). A Computational Model of Referring. In Proceedings of *The Tenth IJCAI*, 640–647. Milan, Italy.

Arens, Y., Hovy, E. & van Mulken, S. (1993a). Structure and Rules in Automated Multimedia Presentation Planning. In Proceedings of *The Thirteenth IJCAI*, volume 2, 1253–1259. Chambéry, France.

Arens, Y., Hovy, E. & Vossers, M. (1993b). Describing the Presentational Knowledge Underlying Multimedia Instruction Manuals. In Maybury, M. (ed.) *Intelligent Multimedia Interfaces*, 280–306. AAAI Press.

Badler, N., Barsky, B. Zeltzer, D. (eds.) (1991a). *Making Them Move: Mechanics, control, and Animation of Articulated Figures*. Morgan Kaufmann: San Mateo, California.

Badler, N., Webber, B., Kalita, J. & Esakov, J. (1991b). *Animation from Instructions. In Badler* et al, 51–93.

Bandyopadhyay, S. (1990). *Towards an Understanding of Coherence in Multimodal Discourse.* Technical Memo TM-90-01, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Saarbrücken, Germany.

Brandt, M., Koch, W., Motsch, W. & Rosengren, I. (1983). Der Einfluß der kommunikativen Strategie auf die Textstruktur – dargestellt am Beispiel des Geschäftsbriefes. In Rosengren, I. (ed.) *Sprache und Pragmatik Lunder Symposium 1982*, 105–135. Almquist & Wiksell: Stockholm.

Costabile, M. F., Catarci, T. & Levialdi, S. (eds.) (1992). *Advanced Visual Interfaces (Proceedings of AVI '92, Rome, Italy).* World Scientific Press: Singapore.

Feiner, S. K. & McKeown, K. R. (1991). Automating the Generation of Coordinated Multimedia Explanations. *IEEE Computer* **24**(10): 33–41.

Grice, H. P. (1975). Logic and Conversation. In Cole, P. & Morgan, J. L. (eds.) *Syntax and Semantics: Speech Acts* **3**: 41–58. Academic Press: New York.

Grimes, J. E. (1975). *The Thread of Discourse.* Mouton: The Hague, Paris.

Hirst, G. (1981). *Anaphora in Natural Language Understanding.* Springer: Berlin, Heidelberg.

Hobbs, J. (1978). *Why is a Discourse Coherent?* Technical Report 176, SRI International: Menlo Park, CA.

Houghton, H. A. & Willows, D. M. (1987). *The Psychology of Illustration, Instructional Issues*, volume 2. Springer: New York, Berlin, Heidelberg, London, Paris, Tokyo.

Hovy, E. H. (1988). Planning Coherent Multisentential Text. In Proceedings of *The Twenty-Sixth ACL*, 163–169.

Hunter, B., Crismore, A. & Pearson, P. D. (1987). Visual Displays in Basal Readers and Social Studies Textbooks. In Willows, D. M. & Houghton, H. A. (eds.) *The Psychology of Illustration, Basic Research*, volume 2, 116–135. Springer: New York, Berlin, Heidelberg.

Kjorup, S. (1978). Pictorial Speech Acts. *Erkenntnis* **12**: 55–71.

Levie, W. H. (1987). Research on Pictures: A Guide to the Literature. In Willows, D. M. & Houghton, H. A. (eds.) *The Psychology of Illustration, Basic Research*, volume 1, 1–50. Springer: New York, Berlin, Heidelberg.

Levin, J. R., Anglin, G. J. & Carney, R. N. (1987). On Empirically Validating Functions of Pictures in Prose. In Willows, D. M. & Houghton, H. A. (eds.) *The Psychology of Illustration, Basic Research* **1**: 51–85. Springer: New York, Berlin, Heidelberg.

Mann, W. C. & Thompson, S. A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization.* Report ISI/RS-87-190. Univ. of Southern California, Marina del Rey, CA.

Marks, J. & Reiter, E. (1990). Avoiding Unwanted Conversational Implicatures in Text and Graphics. In Proceedings of AAAI-90, volume 1, 450–456. Boston, MA.

Maybury, M. (ed.) (1993). *Intelligent Multimedia Interfaces.* AAAI Press.

Molitor, S., Ballstaedt, S.-P. & Mandl, H. (1989). Problems in Knowledge Acquisition from text and Pictures. In Mandl, H. & Levin, J. R. (eds.) *Knowledge Acquisition from text and Pictures*, 3–35. North Holland: Amsterdam, New York, Oxford, Tokyo.

Moore, J. D. & Paris, C. L. (1989). Planning Text for Advisory Dialogues. In Proceedings of *The Twenty-Seventh ACL*, 203–211. Vancouver.

Reiter, E. & Dale, R. (1992). A Fast Algorithm for the Generation of Referring Expressions. In Proceedings of *The Fourteenth COLING*, volume 1, 232–238. Nantes, France.

Roth, S. F., Mattis, J. & Mesnard, X. (1991). Graphics and Natural Language as Components of Automatic Explanation. In Sullivan, J. W. & Tyler, S. W. (eds.) *Intelligent User Interfaces*, 207–239. ACM Press: New York, NY.

Schneiderlöchner, F. (1994). *Generierung von Referenzausdrücken in einem multimodalen Diskurs.* Master's thesis, Fachbereich Informatik, Universität des Saarlandes, Saarbrücken, Germany.

Searle, J. R. (1980). *Speech Acts: An Essay in the Philosophy of Language.* Cambridge University Press: Cambridge, England.

Stock, O. & the ALFRESCO Project Team (1993). ALFRESCO: Enjoying the Combination of Natural Language Processing and Hypermedia for Information Exploration. In Maybury, M. (ed.) *Intelligent Multimedia Interfaces*, 197–224. AAAI Press.

van Dijk, T. A. (1980). *Textwissenschaft.* dtv: München.

Wahlster, W., André, E., Graf, W. & Rist, T. (1991). Designing Illustrated Texts: How Language Production is Influenced by Graphics Generation. In Proceedings of *The Fifth EACL*, 8–14. Berlin, Germany.

Wahlster, W., André, E., Finkler, W. Profitlich, H.-J. & Rist, T. (1993). Plan-Based Integration of Natural Language and Graphics Generation. *AI Journal* **63**: 387–427. Also as DFKI Research Report RR-93-02.

Wazinski, P. (1992). Generating Spatial Descriptions for Cross-Modal References. In Proceedings of *The Third Conference on Applied Natural Language Processing*, 56–63. Trento, Italy.

Willows, D. M. & Houghton, H. A. (1987). *The Psychology of Illustration, Basic Research*, volume 1. Springer: New York, Berlin, Heidelberg, London, Paris, Tokyo.

Wilson, M., Sedlock, D., Binot, J.-L. & Falzon, P. (1992). An Architecture For Multimodal Dialogue. In Proceedings of *The Second Vencona Workshop for Multimodal Dialogue.* Vencona, Italy.