In: A. Ortony and J. Slack and O. Stock (eds.) Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues, Berlin: Springer.

WIP: The Coordinated Generation of Multimodal Presentations from a Common Representation

W. Wahlster, E. André, S. Bandyopadhyay, W. Graf, T. Rist
German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3
W-6600 Saarbrücken 11, Germany
e-mail: {wahlster, andre, graf, rist}@dfki.uni-sb.de

Abstract

The task of the knowledge-based presentation system WIP is the generation of a variety of multimodal documents from an input consisting of a formal description of the communicative intent of a planned presentation. WIP generates illustrated texts that are customized for the intended audience and situation. We present the architecture of WIP and introduce as its major components the presentation planner, the layout manager, the text generator and the graphics generator. An extended notion of coherence for multimodal documents is introduced that can be used to constrain the presentation planning process. The paper focuses on the coordination of contents planning and layout that is necessary to produce a coherent illustrated text. In particular, we discuss layout revisions after contents planning and the influence of layout constraints on text generation. We show that in WIP the design of a multimodal document is viewed as a non-monotonic planning process that includes various revisions of preliminary results in order to achieve a coherent output with an optimal media mix.

1 Introduction

With increases in the amount and sophistication of information that must be communicated to the users of complex technical systems comes a corresponding need to find new ways to present that information flexibly and efficiently. Intelligent presentation systems are important building blocks for the next generation of user interfaces, because they translate from the narrow output channels provided by most of the current application systems into high-bandwidth communications tailored to the individual user. Since, in many situations, information is only presented efficiently through a particular combination of communication modes, the automatic generation of multimodal presentations is one of the tasks of such presentation systems. Multimodal interfaces combining, e.g., natural language and graphics take advantage of both the individual strength of each communication mode and the fact that several modes can be employed in parallel, e.g., in the text-picture combinations of illustrated documents.

As the title of this paper indicates, it is an important goal of this research not simply to merge the verbalization results of a natural language generator and the visualization results of a knowledge-based graphics generator, but to carefully coordinate graphics and text in such a way that they complement each other.

In this paper, we focus on the coordination of contents planning and layout that is necessary to produce a coherent illustrated text. In particular, we discuss layout revisions after contents planning and the influence of layout constraints on contents planning. In a companion paper (see [Wahlster et al. 91]), we describe the influence of graphical constraints on text generation.

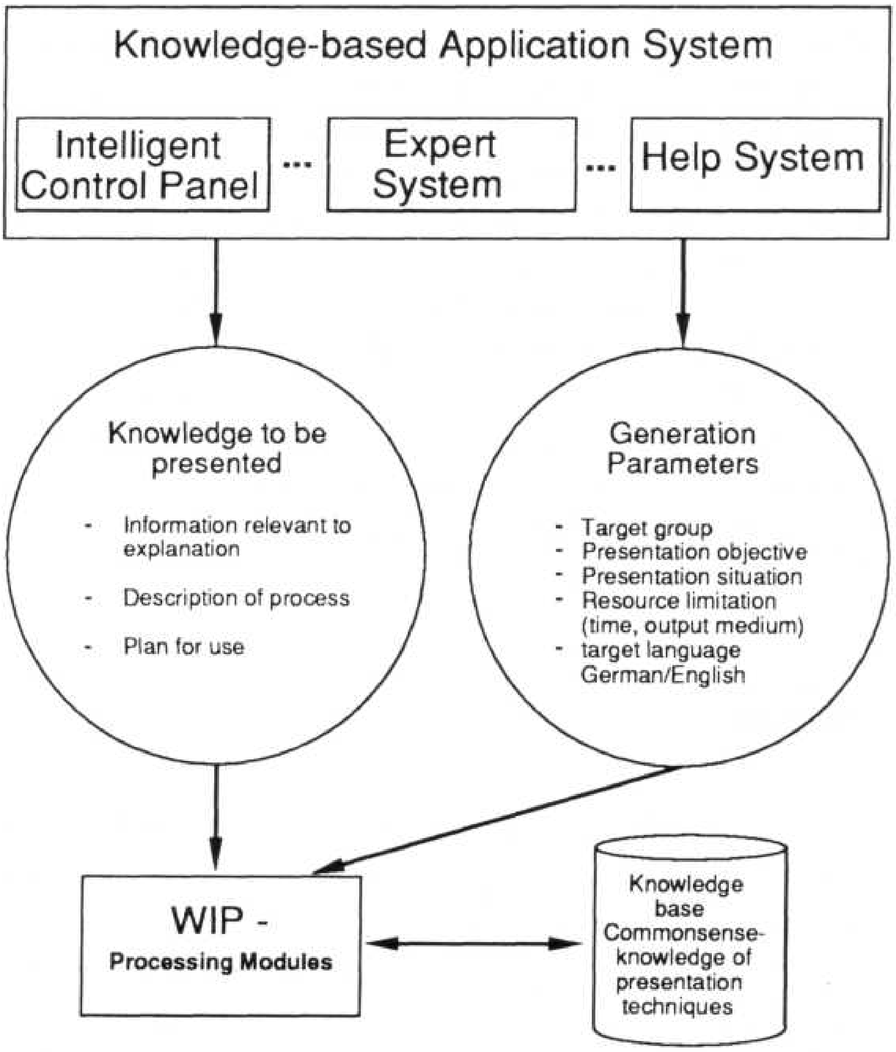

Knowledge-based Application System
Intelligent Control Panel
...
Expert System
...
Help System
Knowledge to be presented
- Information relevant to explanation
- Description of process
- Plan for use
Generation Parameters
- Target group
- Presentation objective
- Presentation situation
- Resource limitation (time, output medium)
- target language German/English
WIP -
Processing Modules
Knowledge base Commonsense-knowledge of presentation techniques

Figure 1: The Generation Parameters of WIP

1.1 WIP: Knowledge-based Presentation of Information

The task of the knowledge-based presentation system WIP is the generation of a variety of multimodal documents from an input consisting of a formal description of the communicative intent of a planned presentation. The generation process is controlled by a set of generation parameters such as target group, presentation objective, resource limitations, and target language (see Fig.1).

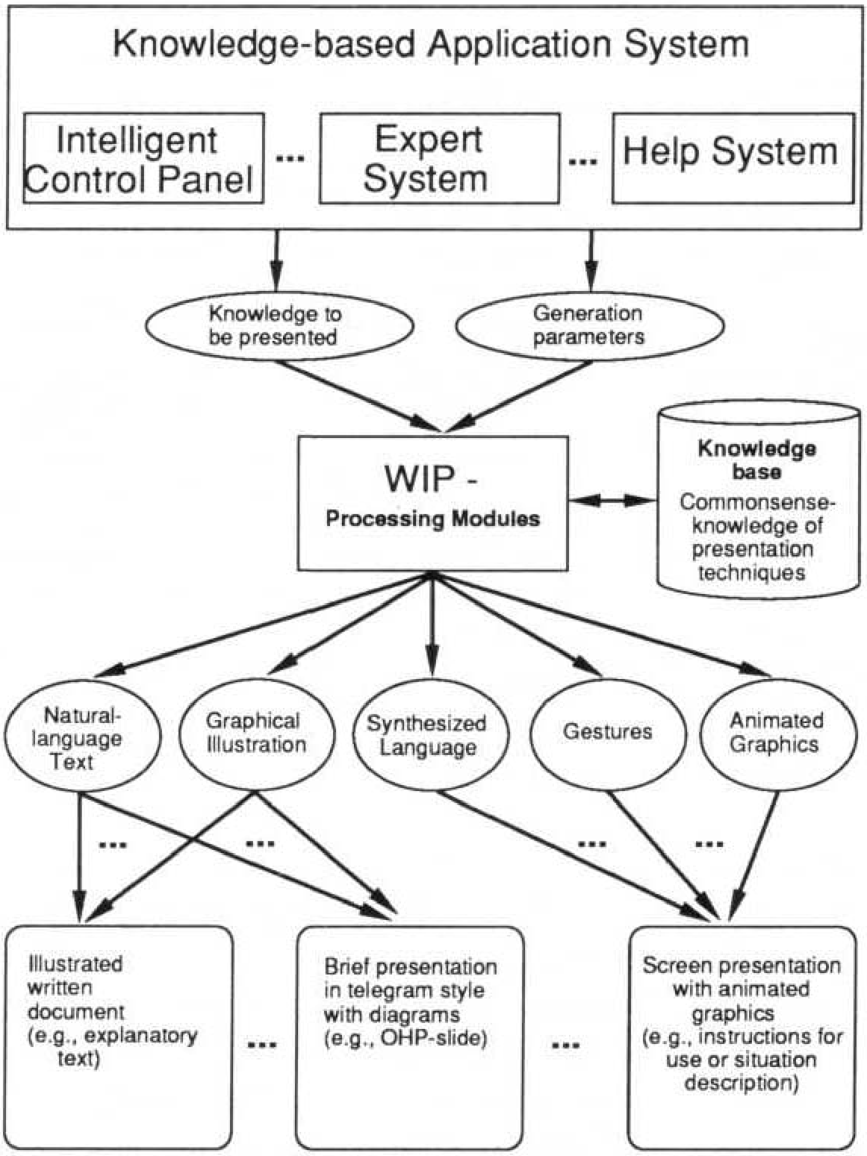

Knowledge-based Application System
Intelligent Control Panel
...
Expert System
...
Help System
Knowledge to be presented
Generation parameters
WIP -
Processing Modules
Knowledge base
Commonsense-knowledge of presentation techniques
Natural-language Text
Graphical Illustration
Synthesized Language
Gestures
Animated Graphics
...
...
...
...
Illustrated written document (e.g., explanatory text)
...
Brief presentation in telegram style with diagrams (e.g., OHP-slide)
...
Screen presentation with animated graphics (e.g., instructions for use or situation description)

Figure2: TheGeneration of a Variety of Multimodal Presentations

This means that the same information content can be presented in a variety of ways depending on the value combination of these generation parameters. Although WIP is designed as a transportable interface to various knowledge-based application systems, such as intelligent control panels, expert systems, and help systems, which supply the presentation system with the necessary input (see Fig. 2), currently all input for the development and testing of the system is created manually.

One of the basic principles underlying the WIP project is that the generation of the various constituents of a multimodal presentation should be generated from a common representation. This leads to the question of how to divide a given communicative goal into subgoals to be realized by the various mode-specific generators, so that they complement each other. This means that we have to explore computational models of the cognitive decision processes coping with questions such as what should go into text, what should go into graphics, and which kinds of links between the verbal and non-verbal fragments are necessary.

A good example of the use of a WIP system is the generation of user-friendly multimodal instructions for technical devices. As a first domain, we have chosen instructions for the use of espresso-machines. Fig. 3 shows a typical text-picture sequence that may be used to instruct a user in filling the watercontainer of an espresso-machine.

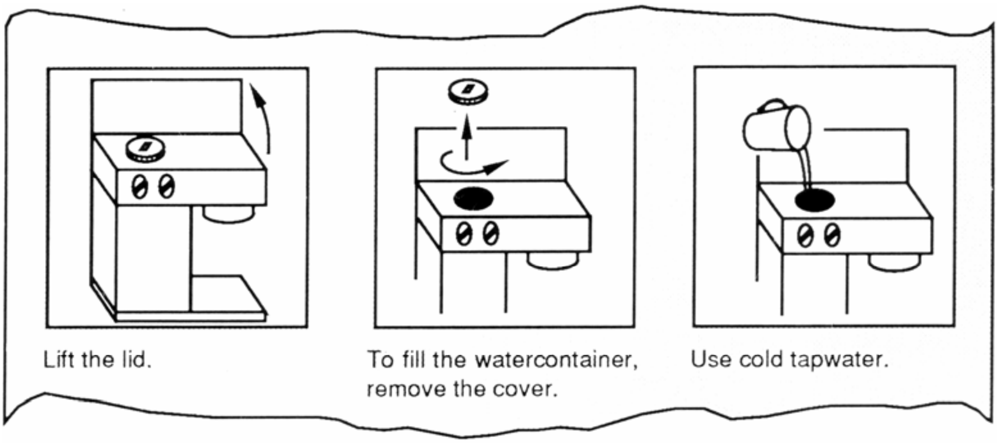


Figure 3: Multimodal Instructions for the Use of an Espresso-Machine

Currently the technical knowledge to be presented by WIP is encoded in a hybrid knowledge representation language of the KL-ONE family including a terminological and assert ional component (see [Nebel 90]). In addition to this propositional representation, which includes the relevant information about the structure, the function, the behavior, and the use of the espresso-machine, WIP has access to an analogical representation of the geometry of the machine in the form of a wire-frame model. This model is used as a basis for the automated design of adequate illustrations.

1.2 Related Research

The automatic design of multimodal presentations has only recently received significant attention in artificial intelligence research. Fig. 4 gives a survey of ongoing projects.

The first group of systems compared in Fig. 4 (XTRA, CUBRICON, ALFresco) consists of multimodal dialog systems with an analysis and generation component. XTRA (cf. [Allgayer et al. 89]) provides multimodal access to an expert system that assists the user in filling out a tax form. CUBRICON (cf. [Neal&Shapiro 88]) is an intelligent interface to

System	Media	Generation of Graphics	Media Coordination	Current Visual Domain	Project Team
XTRA	NL, graphics, pointing	manual	NL and pointing	tax forms	Wahlster et al. (Saarbrücken)
CUBRICON	NL, graphics, pointing	manual	NL and pointing	geographic maps	Shapiro/Neal et al. (Buffalo)
ALFresco	NL, video, pointing	manual	NL and pointing	frescoes	Stock et al. (Trento)
SAGE	NL, graphics	automatic	Not yet	business charts	Roth et al. (CMU)
FN/ANDD	NL, graphics	automatic	Not yet	network diagrams	Marks/Reiter et al. (Harvard)
WIP	NL, graphics	automatic	NL and graphics	espresso machine	Wahlster et al. (Saarbrücken)
COMET	NL, graphics	automatic	NL and graphics	portable radio	Feiner/McKeown et al. (Columbia)

Figure 4: Current Research on Combining Natural Language, Graphics and Pointing

a system for mission planning and situation assessment in a tactical air control domain. ALFresco (cf. [Stock 91]) displays short video sequences about Italian frescoes on a touchscreen and answers questions about details of the videos. In contrast to the first three systems in Fig. 4, the second group currently focuses on the presentation task, although the eventual application environment may also be that of an interactive system.

In the first group of systems, the pointing actions and natural language utterances refer to visual presentations provided by the system builders, whereas the other systems include a component for the generation of graphical displays. All the systems in Fig. 4 combine natural language and graphics, but only systems that generate both forms of presentation from a common representation can address the problem of automatic media choice and coordination. Although both SAGE and FN/ANDD include graphics design components, they have not yet dealt with the problem of media coordination. SAGE creates multimodal explanations of changes in the results generated by quantitative modeling systems (see [Roth et al. 88]). The ANDD (Automated Network-Diagram Designer) system automatically designs network diagrams from a list of relations and a basic network model, whereas the FN system generates natural language expressions describing certain attributes of a particular object shown in the diagrams (see [Marks&Reiter 90]).

The WIP (see [Wahlster et al. 89]) and COMET (see [Feiner&McKeown 89]) projects share a strong research interest in the coordination of text and graphics. They differ from the rest of the systems in that they deal with physical objects (espresso-machine, radio vs. forms, maps, charts, diagrams) that the user can access directly. For example, in the WIP project we assume that the user is looking at a real espresso-machine and uses the presentations generated by WIP to understand the operation of the machine. Likewise COMET generates directions for the maintenance and repair of a portable radio using text coordinated with 3D graphics. In spite of many similarities, there are major differences between COMET and WIP, e.g., in the systems' architecture. While during one of the final processing steps of COMET the media layout component combines text and graphics fragments produced by media-specific generators, in WIP a layout manager interacts with a presentation planner before text and graphics are generated, so that layout considerations can influence the early stages of the planning process and constrain the media-specific generators (see section 3 for more details).

2 The Notion of Coherence for Multimodal Documents

A basic assumption behind the design of WIP is that not only the generation of text, but also the generation of multimodal documents can be considered as a sequence of communicative acts which aim to achieve certain goals (cf. [André&Rist 90a]). As in textlinguistic studies (cf. [Van Dijk 80] and [Mann&Thompson 88]), we distinguish between *main* (MA) and *subsidiary acts* (SA). Main acts convey the kernel of the message. Subsidiary acts serve to support the main acts. In particular, they ensure that necessary preconditions are satisfied, they enhance the effect of the main act or they resolve ambiguities after anticipating the addressee's understanding processes. Since main and subsidiary acts can, in turn, be composed of main and subsidiary acts, we get a hierarchical act structure. 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 61]) or *pictorial acts* (cf. [Kjorup 78]). The structure of a document is, however, not only determined by its act structure, but also by the role acts play in relation to other acts. E.g., one can verbally request an addressee to carry out an action and show with a picture how it should be done. In this example, the act of showing the picture (subsidiary act) is subordinated to the requesting act which conveys the kernel of the message (main act). If the addressee cannot figure out a relation between these acts, the document appears incoherent. Fig. 5 shows a slightly simplified version of the act structure of the instruction sequence in Fig. 3.

Our plan-based approach for the generation of illustrated texts is based on an extended notion of coherence for multimodal documents. In the next sections, we discuss various levels of coherence for picture-sequences and multimodal discourse. The need for coherence constraints the presentation planning process and gives us a criterion for the wellformedness of a complete presentation plan.

Whereas a lot of significant work has been devoted to the study of coherence in text (cf. [Grimes 75], [Hobbs 79], [Hobbs 83], [Reichman 85], [Mann&Thompson 88]), little work has been done in the area of characterising coherence in picture-sequences or in multimodal documents where a segment is either a text segment or a picture, or a combination of both.

In general, coherence can be characterised at three levels: coherence at the *syntactic, semantic* and *pragmatic levels*.

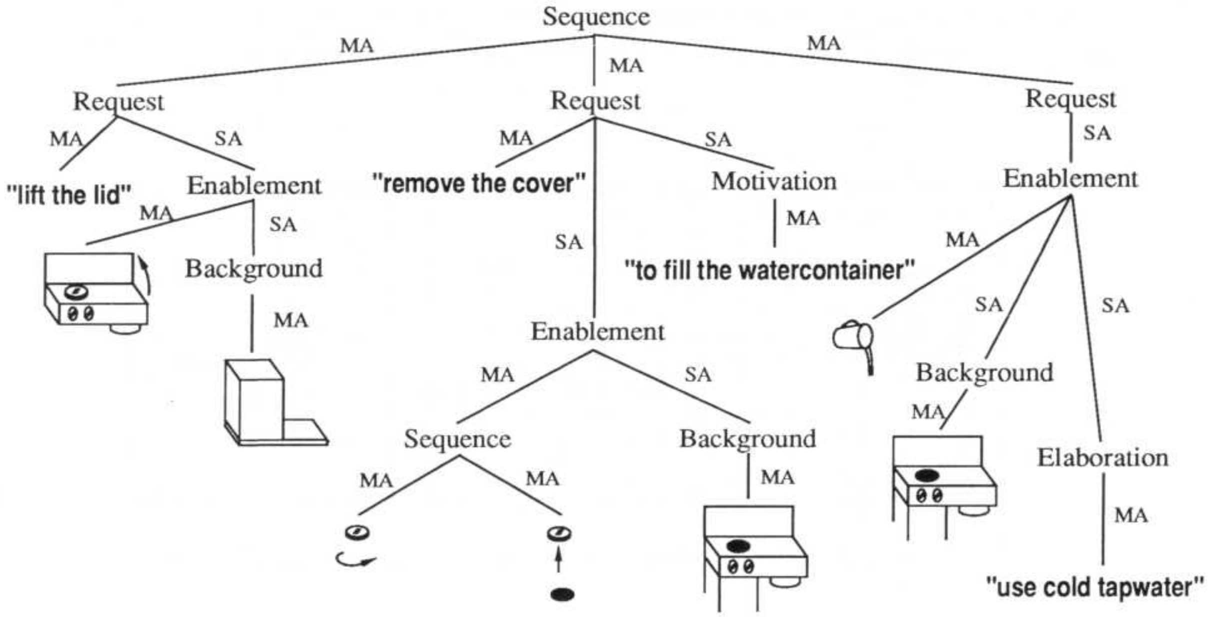


Figure 5: The Action Structure of the Sample Document

Syntactic coherence is a surface-level phenomenon that deals with the immediate connectivity among adjacent segments using some rules or conventions of connectivity. Semantic coherence concerns the content and global structuring of a presentation. It ensures a well-formed thematic organisation of a presentation so that it can be conceived as a unified whole. Pragmatic coherence concerns the effectiveness of a presentation. A presentation is pragmatically coherent to an addressee or a group of addressees if it is compatible with the addressee's interpretive ability (see [Bandyopadhyay 90]).

2.1 Coherence of Picture-Sequences

The syntactic coherence of picture-sequence concerns the immediate connectivity of adjacent pictures. The conventions of the connectivity at the surface level are based on the notion of continuity. We distinguish between the Continuity in perspective (e.g., spatial continuity, continuity in viewpoint and continuity in color), and the continuity of Action (for further details, see [Bandyopadhyay 90]). For example, the picture sequence A-B-C in Fig. 6 illustrating the process of pouring water into the watercontainer of a coffee machine appears to be syntactically incoherent due to the change of perspective from B to C.

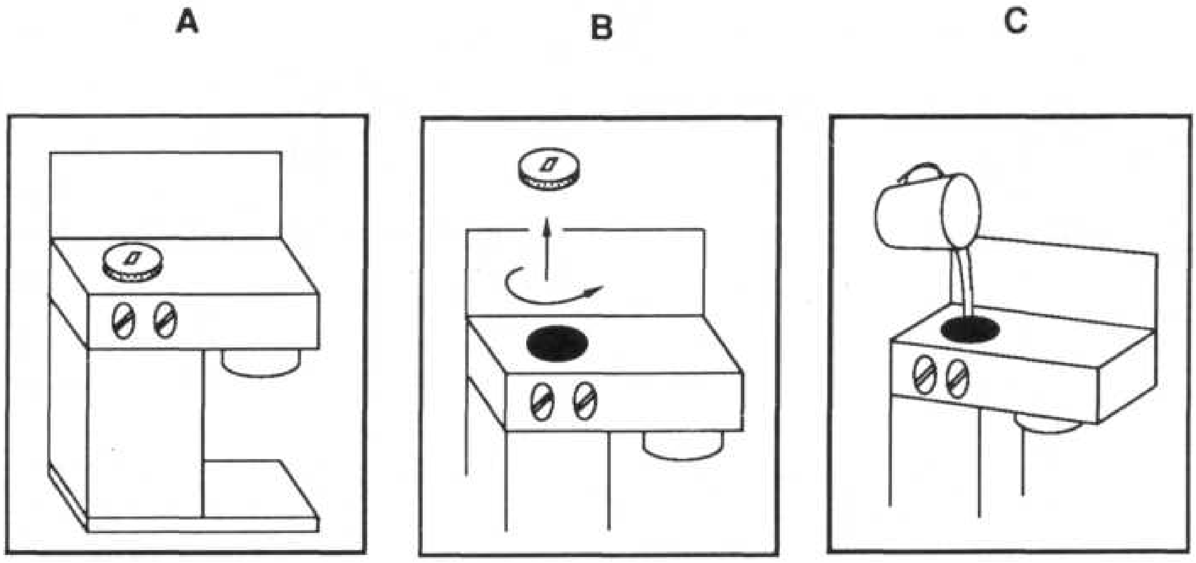


Figure 6: Syntactic Incoherence due to Changed Perspective in C

The discourse structure of a picture-sequence can be described by defining certain coherence relations. These coherence relations are the semantic ties that can exist between two pictures in a sequence, not necessarily adjacent. The picture sequence A-B in Fig. 7 has two possible interpretations:

- The sequence is semantically incoherent due to the unspecified causal relation. Since the default interpretation of a series of pictures showing the same object in different states stipulates a temporal sequence of the snapshots shown, picture A should indicate a cause for the effect of steam shown in picture B.
- The sequence leads the user to the wrong belief that steam appears automatically after some time. In this case, the viewer forces a coherent interpretation by assuming, e.g., a hidden sensor detecting the cup and triggering steam production. This is a typical instance of abductive common-sense reasoning. The picture-sequence leads to an unwanted implicature (cf. [Marks&Reiter 90]), since for the sample machine the user must start the steam production process by switching a knob.

In contrast, the second sequence A'-B7 in Fig. 7 is coherent since the causal relation can be inferred from the change of the switch position. This example shows clearly that for a good design of an illustration the system must find the right level of abstraction. According to Grice's maxim of Relation (cf. [Grice 75]), the graphics designer should avoid irrelevant or spurious graphical elements. The goal of showing the machine in the steam production mode could lead to a sequence like A and B, which shows an abstraction of the machine without the selector switch. To avoid unwanted implicatures, the graphics designer must add more detail. In this case, the extra information showing a change of the switch position in picture-sequence A is necessary.

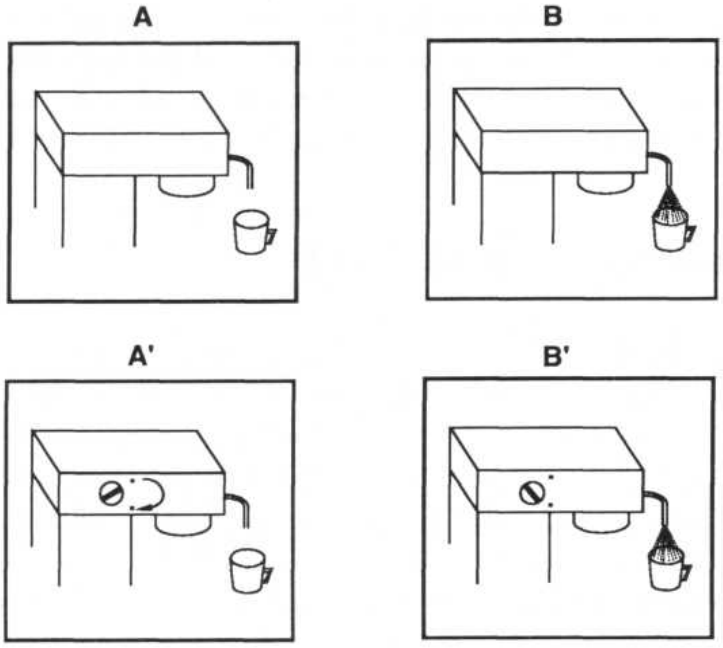


Figure 7: Semantic Incoherence Due to Unspecified Causal Relation

2.2 Coherence of Multimodal Discourse

A linking mechanism between a text segment and a picture segment at the surface level is rather a loose concept, since any text segment can be coupled with any picture at the syntactic level. But one important consideration in the presentation of a multimodal document is the positioning of the picture with respect to the text segment referring to the picture. If the picture is too far away from the relevant text segment or comes after some other pictures, it will lead to surface-level incoherence.

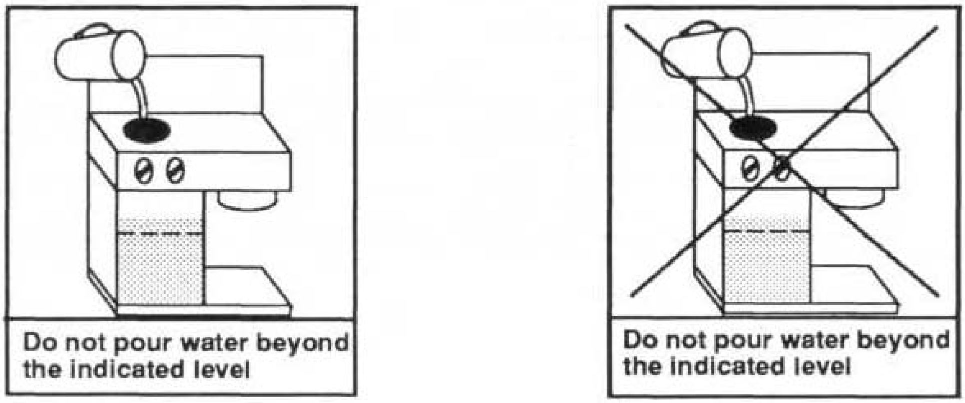

Do not pour water beyond the indicated level
Do not pour water beyond the indicated level

Figure 8: Semantic Incoherence Due to Contradictory Text and Picture Segment

A set of coherence relations can be described to illustrate the semantic tie between a text segment and a picture segment or vice versa (cf. [Bandyopadhyay 90]). Fig. 8 indicates a text-picture combination which is incoherent due to contradictory text and picture segments. Since people tend to skip figure captions for pictures with a straightforward interpretation, they may draw the wrong conclusion: "Pour water beyond the indicated level" The text-picture combination is made coherent by including the negation on a graphical metalevel in the picture. Note that the text is not redundant, but complements the picture because the scope of the negation in the picture is ambiguous.

3 The Architecture of WIP

The architecture of the WIP system guarantees a design process with a large degree of freedom that can be used to tailor the presentation to suit the specific context. During the design process a presentation planner and a layout manager orchestrate the mode-specific generators and the document history handler (see Fig. 9) provides information about intermediate results of the presentation design that is exploited in order to prevent disconcerting or incoherent output. This means that decisions of the language generator may influence graphics generation and that graphical constraints may sometimes force decisions in the language production process.

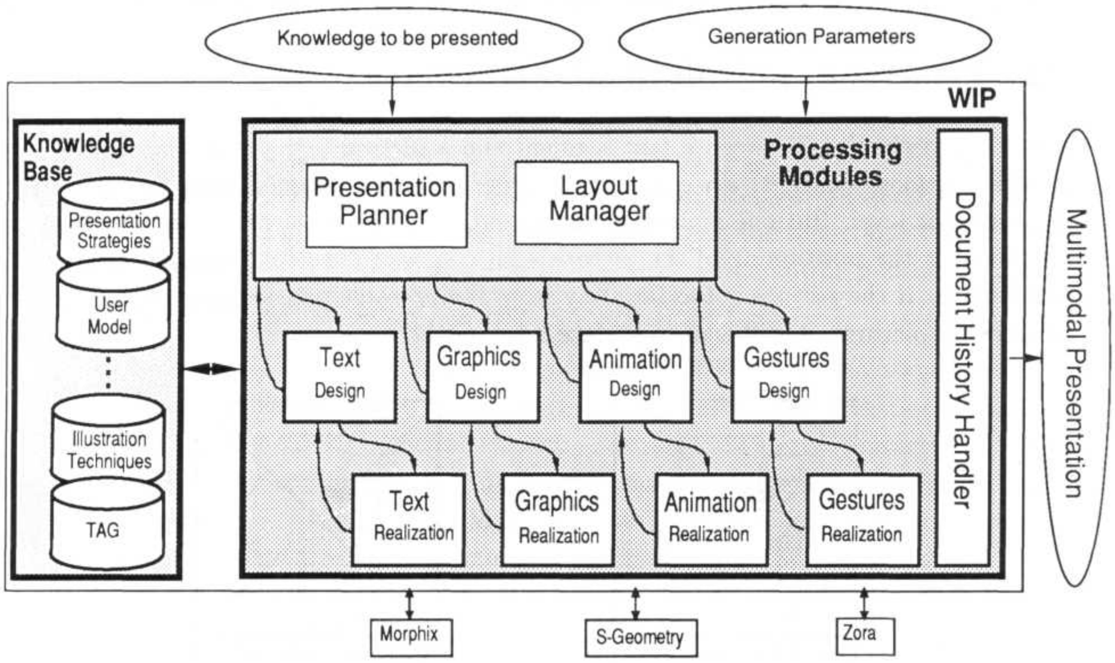


Figure 9: Architecture of the WIP Project

Fig. 9 shows a sketch of WIP's current architecture used for the generation of illustrated documents. Note that WIP includes two parallel processing cascades for the incremental generation of text and graphics. In WIP, the design of a multimodal document is viewed as a non-monotonic process that includes various revisions of preliminary results, massive replanning or plan repairs, and many negotiations between the corresponding design and realization components in order to achieve a fine-grained and optimal division of work between the selected presentation modes.

3.1 The Presentation Planner

The presentation planner is responsible for contents and mode selection. When building up multimodal presentations, one has to know which role a certain document part is to fill and which mode conveys this role most effectively. Currently, we focus on the synthesis of text-picture combinations. Therefore, we have designed presentation strategies that refer to both text and picture production. To represent the strategies, we follow the approach proposed by Moore and colleagues (cf. [Moore&Paris 89] and [Moore&Swartout 89]) to operationalize RST-theory for text planning.

The strategies are represented by a name, a header, an effect, a set of applicability conditions and a specification of main and subsidiary acts. Whereas the header of a strategy indicates which communicative function the corresponding document part is to fill, its effect refers to an intentional goal.[1] The applicability conditions specify when a strategy may be used and put restrictions on the variables to be instantiated. The kernel of the strategies form the main and subsidiary acts. E.g., the strategy below can be used to enable the identification of an object shown in a picture (for further details see [André&Rist 90b]). Whereas graphics should be used to carry out the main act, mode decisions for the subsidiary acts are open.

Name:
```
    Enable-Identification-by-Background
```
Header:
```
    (Provide-Background P A ?x  ?px ?picture GRAPHICS)
```
Effect:
```
    (BMB P A  (Identifiable A ?x ?px ?picture))
```
Applicability Conditions:
```
    (AND  (Bel P  (Perceptually-Accessible A ?x))
          (Bel  P  (Part-of  ?x  ?z)))
```
Main Acts:
```
    (Depict P A (Background ?z) ?pz ?picture)
```
Subsidiary Acts:
```
    (Achieve P (BMB P A (Identifiable A ?z ?pz ?picture)) ?mode)
```

For the automatic generation of illustrated documents, the presentation strategies are treated as operators of a planning system (cf. [André&Rist 90a] and [André&Rist 90b]).

During the planning process, presentation strategies are selected and instantiated according to the presentation task. After the selection of a strategy, the main and subsidiary acts are carried out unless the corresponding presentation goals are already satisfied. Elementary acts, such as 'Depict' or 'Assert', are performed by the text and graphics generators.

[1] In [MooredParis 89], this distinction between header and effect is not made because the effect of their strategies may be an intentional goal as well as a rhetorical relation

3.2 The Layout Manager

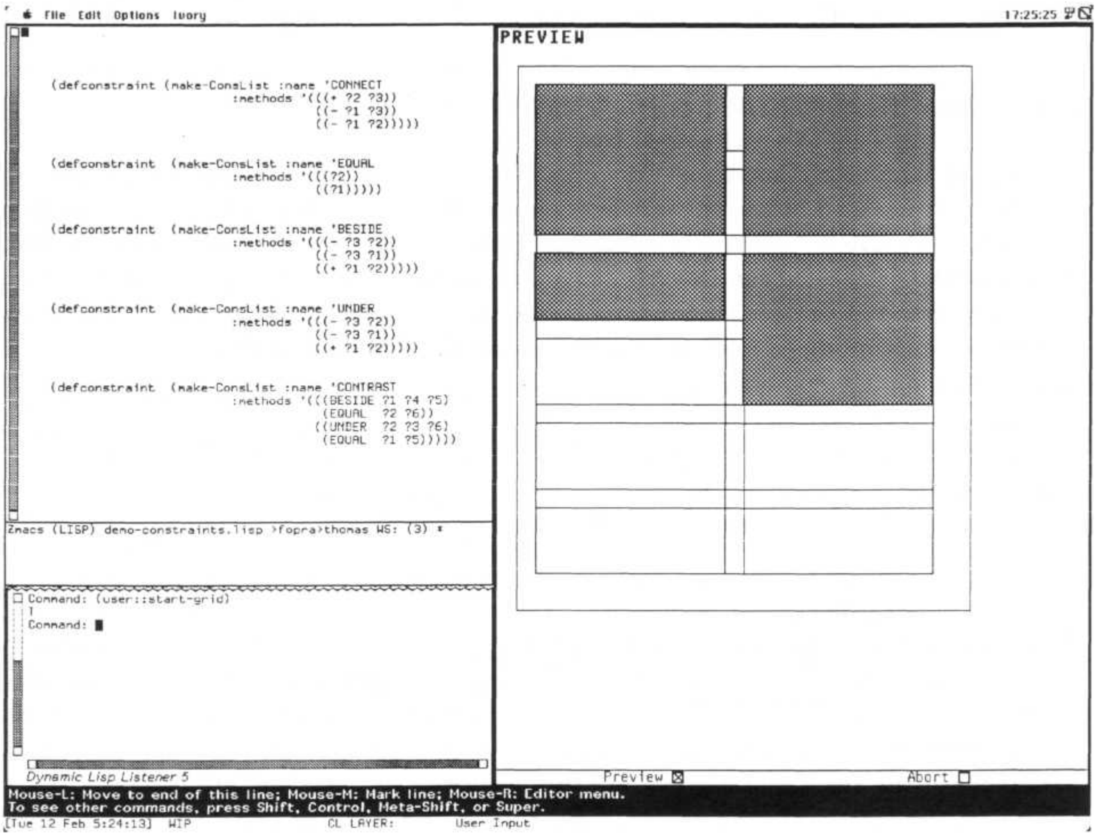
The main task of the layout manager is to convey certain semantic and pragmatic relations specified by the planner by the arrangement of graphic and text fragments received from the mode-specific generators, i.e., to determine the size of the boxes and the exact coordinates for positioning them on the document page. Therefore, we use a grid-based approach as an ordering system for efficiently designing functional (i.e., uniform, coherent, and consistent) layouts (cf. [Müller-Brockmann 81]). This method is also used by Beach for low-level table layout (cf. [Beach 85]) and in the GRID system for automating display layout (cf. [Feiner 88]).

The layout process is carried out in two phases with different levels of detail. In the first phase, a draft version of a high-level page layout is produced. Since at that stage of the process neither the text generator nor the graphics generator has produced any output, the layout manager only has information about the contents, the act structure and the selected mode combination which is available via the document history handler. Thus, the layout manager uses default assumptions to determine a skeletal version of an initial page layout based on uninstantiated text and graphic boxes. As soon as a generator has supplied any output, the corresponding box is instantiated and the incremental process of low-level layout planning can start. Then the layout manager has to position this box on the grid considering design restrictions. As the example below shows, design constraints or visual unbalances in the output presentation can require a total revision of the skeletal layout or in the worst-case even a change of the contents.

A central problem when automatically designing layout is the representation of design-relevant knowledge. According to [Borning&Duisberg 86], constraint networks seem to be a natural formalism to declaratively incorporate aesthetic knowledge into the geometric layout process. Layout constraints can be classified as semantic, geometric and topological, and temporal. Semantic constraints essentially correspond to coherence relations, such as sequence and contrast, and can be easily reflected through specific design constraints. They describe perceptual criteria concerning the organization of the boxes, such as the sequential ordering (horizontal or vertical layout), alignment, grouping, symmetry or similarity.

When using constraints to represent layout knowledge, one often wants to prioritize the constraints in those which must be required and others which are preferably held. A powerful way of expressing this layout feature is to organize the constraints in a hierarchy by assigning a preference scale to the constraint network. We distinguish obligatory, optional and default constraints. The latter state default values, which remain fixed unless the corresponding constraint is removed by a stronger one. Since there are constraints that only have local effects, the constraint hierarchy has to be changed frequently. The constraint solver must therefore be able to add and remove constraints dynamically during runtime.[2]

[2] A theory of constraint hierarchies is described in [Borning et al. 89]. An incremental constraint hierarchy solver (cf. also the DeltaBlue algorithm [Freeman-Benson 90]) for WIP has been implemented by Wolfgang MaaB (cf. [MaaB 91])



```
(defconstraint (make-ConsList :name 'CONNECT
                              :methods '(((+ ?2 ?3))
                                         ((- ?1 ?3))
                                         ((- ?1 ?2))))

(defconstraint  (make-ConsList :name 'EQUAL
                              :methods '(((?2))
                                         ((?1)))))

(defconstraint  (make-ConsList :name 'BESIDE
                              :methods '(((- ?3 ?2))
                                         ((- ?3 ?1))
                                         ((+ ?1 ?2)))))

(defconstraint  (make-ConsList :name 'UNDER
                              :methods '(((- ?3 ?2))
                                         ((- ?3 ?1))
                                         ((+ ?1 ?2)))))

(defconstraint  (make-ConsList :name 'CONTRAST
                              :methods '(((BESIDE ?1 ?4 ?5)
                                          (EQUAL  ?2 ?6))
                                         ((UNDER  ?2 ?3 ?6)
                                          (EQUAL  ?1 ?5)))))
```

```
Zmacs (LISP) demo-constraints.lisp >fopra>thomas WS: (3) *
```

```
□ Command: (user::start-grid)
  T
  Command: ■
```

```
Mouse-L: Move to end of this line; Mouse-M: Mark line; Mouse-R: Editor menu.
To see other commands, press Shift, Control, Meta-Shift, or Super.
[Tue 12 Feb 5:24:13]  WIP            CL LAYER:         User Input
```

Figure 10: Constraint definition and a preview showing a grid populated with two contrasting graphic boxes and the corresponding textboxes.

A typical example of using a constraint hierarchy in geometric layout is the problem of leaving enough white space between two graphic boxes communicating a contrast. The adequate aesthetic criteria can be represented by three constraints of different strength: one obligatory constraint that specifies that the distance between the boxes must be greater than zero and a disjunction of two optional constraints that the boxes are preferably aligned side by side or else below each other. To give an example of a typical compound constraint in the syntax of a constraint language, let's have a look at a section of the definition of the 'contrast'-constraint (cf. Fig. 10). Since the ordering of the constraints in the definition is significant, the stronger constraints should precede the weaker ones. E.g., according to the definition above, the layout manager will use a horizontal alignment in preference to a vertical one if a contrast-constraint has to be satisfied. For a detailed description of the layout manager see [Graf 90].

3.3 The Text Generator

WIP's text generator is based on the formalism of tree adjoining grammars (TAGs^. In particular, lexicalized TAGs with unification are used for the incremental verbalization of logical forms produced by the presentation planner (cf. [Harbusch 90], [Schauder 90]). The grammar is divided into an LD (local dominance) and an LP (linear precedence) part so that the piecewise construction of syntactic constituents is separated from their linearization according to word order rules (cf. [Finkler&Neumann 89]).

The text generator uses a TAG parser in a local anticipation feedback loop (see [Jameson&Wahlster 82]). The generator and parser form a bidirectional system, i.e., both processes are based on the same TAG. By parsing a planned utterance, the generator makes sure that it does not contain unintended structural ambiguities.

Since the TAG-based generator is used in designing illustrated documents, it has to generate not only complete sentences, but also sentence fragments such as NPs, PPs, or VPs, e.g., for figure captions, section headings, picture annotations, or itemized lists. Given that capability and the incrementality of the generation process, it becomes possible to interleave generation with parsing in order to check for ambiguities as soon as possible. Currently, we are exploring different domains of locality for such feedback loops and trying to relate them to resource limitations specified in WIP's generation parameters. One parameter of the generation process in the current implementation is the number of ad-joinings allowed in a sentence. This parameter can be used by the presentation planner to control the syntactic complexity of the generated utterances and sentence length. If the number of allowed adjoinings is small, a logical form that can be verbalized as a single complex sentence may lead to a sequence of simple sentences. The leeway created by this parameter can be exploited for mode coordination. For example, constraints set up by the graphics generator or layout manager can force delimitation of sentences, since in a good design, picture breaks should correspond to sentence breaks, and vice versa (see [McKeown&Feiner 90]).

3.4 The Graphics Generator

When generating illustrations of physical objects WIP does not rely on previously authored picture fragments or predefined icons stored in the knowledge base. Rather, we start from a hybrid object representation that includes a wireframe model for each object. Although these wireframe models, along with a specification of physical attributes, such as surface color or transparency, form the basic input of the graphics generator, the design of illustrations is regarded as a knowledge-intensive process that exploits various knowledge sources to achieve a given presentation goal efficiently. E.g., when a picture of an object is requested, we have to determine an appropriate perspective in a context-sensitive way (cf. [Rist& André 90]). In our approach, we distinguish between three basic types of graphical techniques. First, there are techniques to create and manipulate a 3D object configuration that serves as the subject of the picture. E.g., we have developed a technique to spatially separate the parts of an object in order to construct an exploded view. Second, we can choose among several techniques that map the 3D subject onto its depiction. E.g., we can construct either a schematic line drawing or a more realistic looking picture using rendering techniques. The third kind of technique operates on the picture level. E.g., an object depiction may be annotated with a label (see Fig. 11), or picture parts may be colored in order to emphasize them. The task of the graphics designer is then to select and combine these graphical techniques according to the presentation goal. The result is a so-called design plan which can be transformed into executable instructions of the graphics realization component. This component relies on the 3D graphics package S-Geometry and the 2D graphics software of the Symbolics window system.

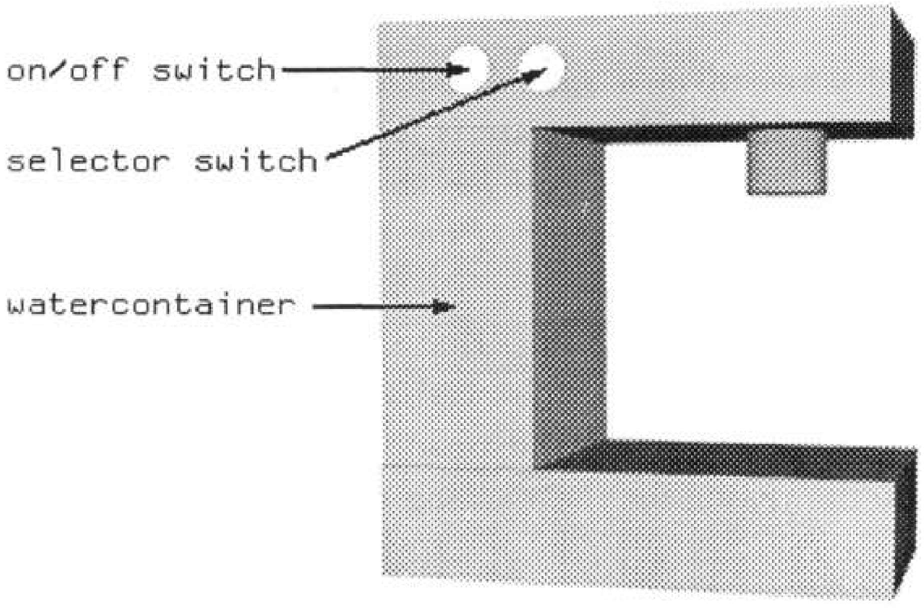


Figure 11: Rendered Picture with Annotations

3.5 Tailoring Presentations to the Addressee

One advantage of the automated design of multimodal documents over the display of predefined presentations, e.g., in conventional hypermedia systems, is that in a knowledge-based presentation system like WIP the generated document can be tailored to a particular target group and presentation situation. As mentioned in section 1, one of the generation parameters of WIP is information about each individual target group or addressee. If the generated multimodal document is to be informative, understandable and effective in reaching the presentation goal specified in the input, the presentation system has to take into account factors like the addressee's prior knowledge about the domain and his level of expertise, i.e., the system has to exploit a user model (cf. [Wahlster&Kobsa 89]).

The user modeling component of WIP provides the presentation planner with information about the addressee that affects the content and structure of the generated document.

Let's discuss how WIP can use the assumptions about an addressee's domain knowledge contained in the user model to tailor the presentation to each addressee. Suppose that the system's present task is to generate a warning against opening the cover of the water-container too early after having used the espresso machine. If the system assumes that the addressee has no detailed knowledge about the preparation of espresso, some motivation should procede the warning itself. In our example, the extreme pressure and high temperature in the watercontainer are the main reasons for the warning. If the system assumes that the addressee does not know the reasons for the extreme pressure and high temperature, it should introduce them before the warning.

In the presentation situation just described, a text like (1) would be communicatively adequate.

(1) *Espresso is coffee prepared in a special machine from finely ground coffee beans, through which steam under high pressure is forced. Because of the extreme pressure and high temperature, you should wait for at least two minutes after switching off the machine before you open the cover of the watercontainer.*

In the opposite case, when the system assumes that the addressee has already used another type of espresso machine, the system can just verbalize a warning like (2). Note that (2) would be pragmatically incoherent (cf. [Bandyopadhyay 90]) for the first type of addressee introduced above, since the reason for the warning would remain unclear to him.

(2) *Wait at least for two minutes after switching off the machine before you open the cover of the watercontainer.*

It is obvious that WIP's user model should not only constrain the text planning, but also guide other processes like media choice, gesture generation (see [Wahlster 91]), and the synthesis of graphics.

4 Coordination of Contents Planning and Layout

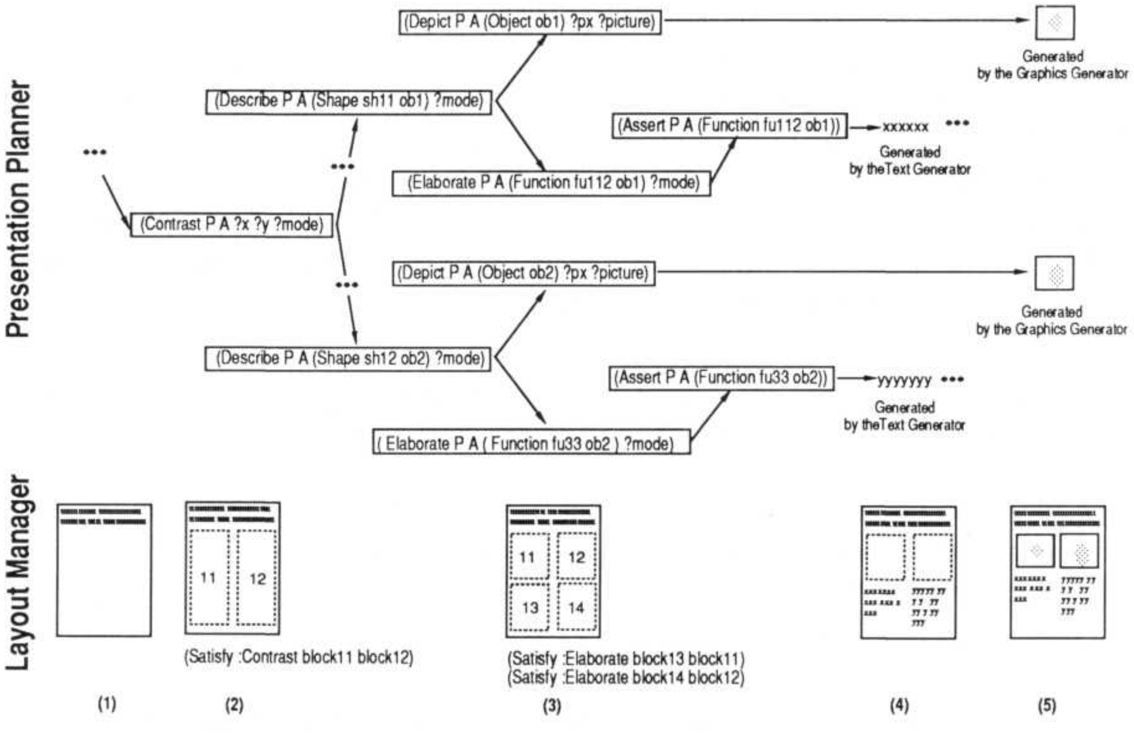

Presentation Planner
(Depict P A (Object ob1) ?px ?picture)
Generated
by the Graphics Generator
(Describe P A (Shape sh11 ob1) ?mode)
(Assert P A (Function fu112 ob1))
xxxxxx
Generated
by theText Generator
(Elaborate P A (Function fu112 ob1) ?mode)
(Contrast P A ?x ?y ?mode)
(Depict P A (Object ob2) ?px ?picture)
Generated
by the Graphics Generator
(Describe P A (Shape sh12 ob2) ?mode)
(Assert P A (Function fu33 ob2))
yyyyyyy
Generated
by theText Generator
(Elaborate P A (Function fu33 ob2) ?mode)
Layout Manager
11
12
(Satisfy :Contrast block11 block12)
11
12
13
14
(Satisfy :Elaborate block13 block11)
(Satisfy :Elaborate block14 block12)
(1)
(2)
(3)
(4)
(5)

Figure 12: The Coordination of Content Planning and Layout

To illustrate the temporal coordination of content planning and layout, some snapshots of the processes are shown in Fig. 12. Suppose that the initial layout consists of an instantiated text block on the top of the page (stage 1). Let's assume the planner has decided to compare two objects obi and ob2. To highlight the contrast-relationship between the planned document parts, two default boxes are placed side by side (stage 2).

After the plan has been refined, the layout manager knows that the contrast between the two objects will be communicated through two pictures and two text boxes. Note that in this processing phase neither the text generator nor the graphics generator has been activated. Thus, the size of the boxes in the initial layout is determined by default values computed from the presentation plan generated so far. The two explanatory text fragments are placed in two columns aligned with the corresponding graphics boxes in order to emphasize the comparison (note that exchanging the text fragments in both columns would result in an incoherent text-picture combination (stage 3). As shown in Fig. 9, WIP's architecture contains two parallel processing cascades for the generation of text and graphics. At stage 4 in the figure, the text generator has already produced a first version of the two paragraphs, whereas the graphics generator is not yet ready. Thus, the layout manager instantiates the corresponding boxes. Finally, the picture boxes are filled (stage 5).

4.1 Revising Layout after Contents Planning

Frequently, a draft layout has to be revised because the output supplied by the generators does not fit into the planned boxes. When a partially instantiated layout entered in the document history is evaluated by the layout manager with a negative result, a dependency-based layout revision process is initiated.

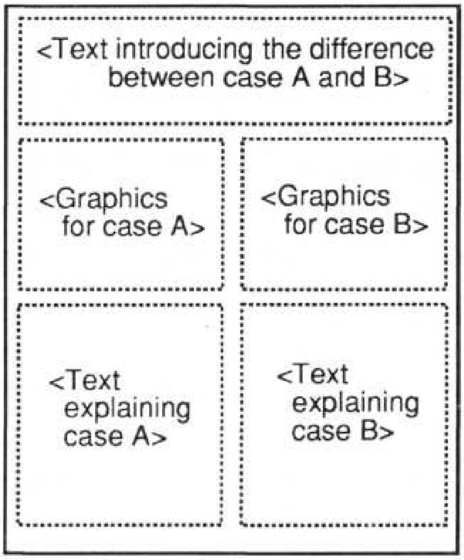


Figure 13: Planned Layout Skeleton

Let's assume as in the example above, that the presentation planner has decided to describe the difference between two concepts A and B (e.g., the preparation of espresso or cappuchino in our domain) by three basic blocks: a paragraph introducing the difference, two figures illustrating the difference, and two verbal explanations of other distinguishing features of A and B, which are not shown in the graphics. Starting from this information, the layout manager produces a skeletal version of a preferred page layout that consists of five boxes (two for graphics and three for text) placed on a grid (see Fig. 13).

Suppose that it turns out during the text generation process in the example above that the distinguishing features of A can be explained much less verbosely than those of B (cf. Fig. 14a). As a consequence, the text columns A and B would become completely unbalanced. In an extreme case, when the text fragment on B does not fit on the current page, the picture-text combination even can become syntactically incoherent (see section 2.2), since the rest of the information on B is presented in the first lines of the next page of the generated document.

The revised layout (see Fig. 14b) again pairs the corresponding graphics and text blocks, but does not contrast them directly by placing them side by side. Although the text for B does not fit on the page since some space is lost by separating and centering the graphics blocks, the resulting illustrated document is coherent.

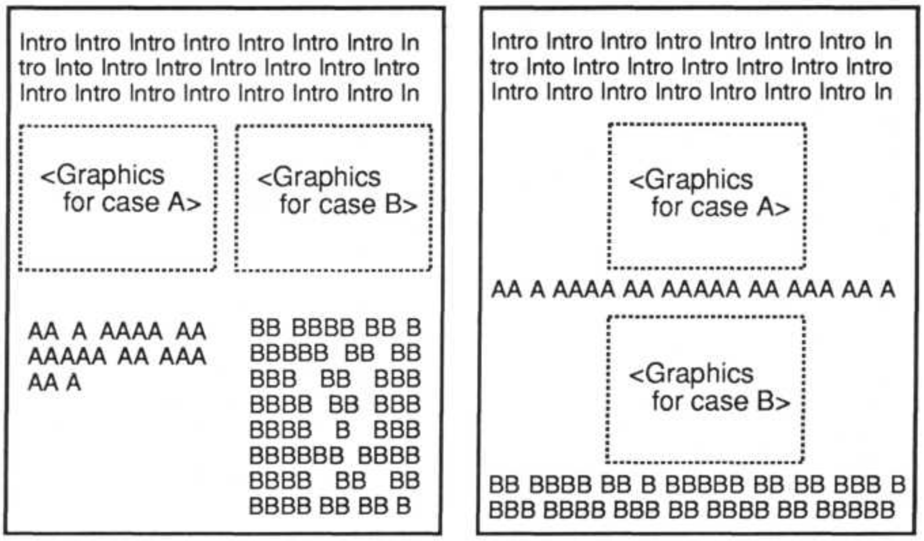

Intro Intro Intro Intro Intro Intro Intro In
tro Into Intro Intro Intro Intro Intro Intro
Intro Intro Intro Intro Intro Intro Intro In
<Graphics for case A>
<Graphics for case B>
AA A AAAA AA AAAAA AA AAA AA A
BB BBBB BB B BBBBB BB BB BBB BB BBB BBBB BB BBB BBBB B BBB BBBBBB BBBB BBBB BB BB BBBB BB BB B
Intro Intro Intro Intro Intro Intro Intro In
tro Into Intro Intro Intro Intro Intro Intro
Intro Intro Intro Intro Intro Intro Intro In
<Graphics for case A>
AA A AAAA AA AAAAA AA AAA AA A
<Graphics for case B>
BB BBBB BB B BBBBB BB BB BBB B
BBB BBBB BBB BB BBBB BB BBBBB

Figure 14: (a) Partially Instantiated Layout, (b) Revised Layout

4.2 Revising Contents after Layout

There are also cases in which formatting restrictions influence the selection of the contents. Such restrictions may be given a priori (e.g., when a certain format is required) or result during the generation process (e.g., when the system has to follow the format of previously generated document parts to ensure syntactic coherence).

To illustrate such a situation, let's assume that the presentation goal is to request the addressee to lift the lid of the watercontainer. The planner decides to convey the actual request through text and to show in a picture how the requested action should be carried out. Since the planner is not sure whether the addressee knows why the action should be carried out, it decides to mention the purpose of the action as a motivation.

The layout manager generates a draft layout consisting of a picture and a text box. Let's suppose that the size of the boxes is determined by the size of previously generated text and picture boxes.

After text production, the layout manager discovers that the generated text exceeds the box lines (cf. Fig. 15a). Due to the severe format restrictions, it has no chance to increase the text box. Therefore, the layout manager sends a message to the text generator to shorten the sentence. Since the text generator is not able to produce significantly shorter paraphrases by ellipses, pronouns, etc., and is not allowed to manipulate the contents specification, it informs the presentation planner that the required task cannot be accomplished.

The presentation planner then evaluates which contents reduction will have the least effects on the success of the communication process. Since the main message to be conveyed is to request the addressee to open the lid, it decides to leave out the motivation. The text generator is now able to communicate the message through a sentence that fits into lie box (cf. Fig. 15b).

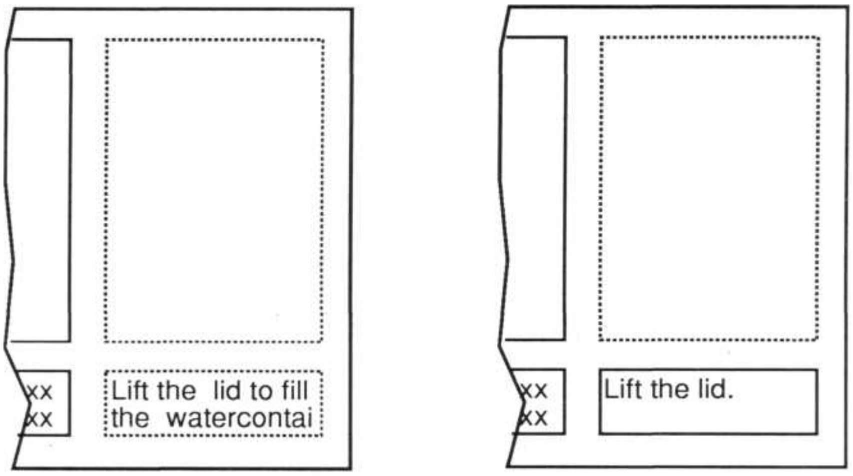


Figure 15: (a) Planned Layout Skeleton, (b) Revised Layout

5 Conclusions

In this paper, we presented a computational model for the generation of multimodal communications. We showed how the knowledge-based presentation system WIP coordinates graphics and text in such a way that they complement each other in an illustrated document. The basic principles underlying the WIP project are that the generation of all constituents of a multimodal presentation should start from a common representation and that the design of a text-picture sequence can be modeled as a non-monotonic planning process. We showed how WIP's presentation planner and layout manager orchestrate the text and graphics generator during the design process. An extended notion of coherence1 for multimodal documents was introduced that is used to constrain the presentation planning process. The paper focused on the coordination of contents planning and layout that is necessary to produce a coherent illustrated text. In particular, we discussed layout revisions after contents planning and the influence of layout constraints on text generation.

6 The Current Status of the WIP Project

The WIP project is supported by the German Ministry of Research and Technology under grant ITW8901 8 and was started in May 1989 for a 4-year period. The project team is headed by Wolfgang Wahlster and is divided into three subgroups for presentation planning, language generation and knowledge representation. The presentation planning group focuses on problems of context-directed selection of contents, automated graphics design, coordination of text and graphics (Elisabeth André and Thomas Rist), and constraint-based layout (Winfried Graf). The language generation group works on the incremental and parallel generation of text using lexicalized tree-adjoining grammars with feature unification (Karin Harbusch, Anne Schauder and Wolfgang Finkler). The knowledge representation group focuses on extending the expressiveness of the terminological logic used in WIP with regard to the representation of temporal relations, action structures, default values and exceptions (Bernhard Nebel, Jochen Heinsohn and Hans-Jürgen Profitlich). Testbed modules for the various components of the WIP system are currently being implemented on Maclvory systems in Common Lisp/CLOS.

The development of WIP is an ongoing group effort and has benefited from the contributions of our students Andreas Butz, Bernd Hermann, Antonio Krüger, Daniel Kudenko, Wolfgang Maafi, Thomas Schiffmann, Georg Schneider, Frank Schneiderlöchner, Christoph Schommer, Dudung Soetopo, Peter Wazinski, and Detlev Zimmermann.

References

[Allgayer et al. 89] J. Allgayer, K. Harbusch, A. Kobsa, C. Reddig, N. Reithinger, D. Schmauks. XTRA: A Natural Language Access System to Expert Systems. In: International Journal of Man-Machine Studies Vol. 31, pp. 161-195, 1989.

[André&Rist 90a] E. André, T. Rist. Towards a Plan-Based Synthesis of Illustrated Documents. In: Proc. of the 9th European Conference on Artificial Intelligence, pp. 25-30, 1990.

[André&Rist 90b] E. André, T. Rist. Synthesizing Illustrated Documents: A Plan-Based Approach. In: Proc. of InfoJapan 90, Vol. 2, pp. 163-170, 1990.

[Bandyopadhyay 90] S. Bandyopadhyay. Towards an Understanding of Coherence in Multimodal Discourse. Technical Memo DFKI-TM-90-01, German Research Center for Artificial Intelligence, Saarbrücken, 1990.

[Beach 85] R. J. Beach. Setting Tables and Illustrations with Style. Xerox PARC Technical Report CSL-85-3, 1985.

[Borning&Duisberg 86] A. Borning and R. Duisberg. Constraint-Based Tools for Building User Interfaces. ACM Trans, on Graphics 5:6, 345-374, 1986.

[Borning et al. 89] A. Borning, B. Freeman-Benson, and M. Wilson. Constraint Hierarchies. Internal Report, Department of Computer Science and Engineering, FR- 35, University of Washington, Seattle, 1989.

[Feiner 88] S. Feiner. A Grid-Based Approach to Automating Display Layout. In: Proc. Graphics Interface 88. Palo Alto: Morgan Kaufmann, pp. 192-197, 1988.

[Feiner&McKeown 89] S. Feiner and K. McKeown. Coordinating Text and Graphics in Explanation Generation. In: DARPA Speech and Natural Language Workshop, 1989.

[Finkler&Neumann 89] W. Finkler and G. Neumann. POPEL-HOW: A Distributed Parallel Model for Incremental Natural Language Production with Feedback. In: Proc. of the 11th IJCAI, pp. 1518-1523, 1989.

[Freeman-Benson et al. 90] B. Freeman-Benson, J. Maloney, and A. Borning. An Incremental Constraint Solver. Communications of the ACM, Vol. 33, No. 1, pp. 54- 63, 1990.

[Graf 90] W. Graf. Spezielle Aspekte des automatischen Layout-Designs bei der koordinierten Generierung von multimodalen Dokumenten. GI-Workshop "Multimediale elcktronische Dokumente", 1990.

[Grice 75] H. Grice. Logic and Conversation. In: Cole and Morgan (Eds.). Syntax and Semantics, Vol. 3, New York: Academic Press, 1975.

[Grimes 75] J.E. Grimes. The Thread of Discourse. Mouton: The Hague, Paris, 1975.

[Harbusch 90] K. Harbusch. Constraining Tree Adjoining Grammars by Unification. Proc. of the 13th COLING, pp. 167-172, 1990.

[Ilobbs 79] J. Hobbs. Coherence and Coreference. Cognitive Science 3(1), 1979.

[Hobbs 83] J. Hobbs. Why is Discourse Coherent? In: Neubauer (ed.). Coherence in Natural Language Texts. Hamburg: Buske, 1983.

[Jamrson&Wahlster 82] A. Jameson and W. Wahlster. User Modelling in Anaphora Generation: Ellipsis and Definite Description. In: Proc. of the 5th ECAI, pp. 222-227, 1982.

[Kjorup 78] S. Kjorup. Pictorial Speech Acts. In: Erkenntnis 12, pp. 55-71, 1978.

[MannfcThompson 88] W. Mann and S. Thompson. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. In: TEXT, 8(3), 1988.

[Marks&Reiter 90] J. Marks and E. Reiter. Avoiding Unwanted Conversational Implicatures in Text and Graphics. In: Proc. of the 8th AAAI, pp. 450-455, 1990.

[Maafi 91] W. MaaB. Constraint-basierte Representation von graphischem Wissen am Beispiel des Layout-Managers in WIP. MS thesis, Computer Science, University of Saarbrücken, 1991.

[McKeown&Feiner 90] K. McKeown and S. Feiner. Interactive Multimedia Explanation for Equipment Maintenance and Repair. In: DARPA Speech and Natural Language Workshop, pp. 42-47, 1990.

[Moore&Paris 89] J. Moore and C. Paris. Planning Text for Advisory Dialogues. In: Proc. of the 27th ACL, pp. 203-211, 1989.

[Moore&Swartout 89] J.D. Moore and W.R. Swartout. A Reactive Approach to Explanation. In: Proc. of the 11th IJCAI, pp. 1504-1510, 1989.

[Muller-Brockmann 81] J. Müller-Brockmann. Grid Systems in Graphic Design. Stuttgart: Ilatje, 1981.

[Neal&Shapiro 88] J. Neal and S. Shapiro. Intelligent Multi-Media Interface Technology. In: Proc. of the Workshop on Architectures of Intelligent Interfaces: Elements&Prototypes, pp. 69-91, 1988.

[Nebel 90] B. Nebel. Reasoning and Revision in Hybrid Representation Systems. Lecture Notes in AI, Vol. 422, Berlin: Springer-Verlag, 1990.

[Reichmann 85] R. Reichmann. Getting Computers to Talk like You and Me. Cambridge, MA: MIT Press, 1985.

[Rist&André 90] T. Rist and E. André. Wissensbasierte Perspektivenwahl fur die automatische Erzeugung von 3D-Objektdarstellungen. In: K. Kansy and P. Wifikirchen (Eds.). Graphik und KI. IFB 239, Berlin: Springer-Verlag, pp. 48-57, 1990.

[Roth et al. 88] S. Roth, J. Mattis, and X. Mesnard. Graphics and Natural Language as Components of Automatic Explanation. In: Proc. of the Workshop on Architectures of Intelligent Interfaces: Elements&Prototypes, pp. 109-128, 1988.

[Searle 69] J. Searle. Speech Acts: An Essay in the Philosophy of Language. Cambridge, MA: Cambridge University Press, 1969.

[Schauder 90] A. Schauder. Inkrementelle syntaktische Generierung natürlicher Sprache mit Tree Adjoining Grammars. MS thesis, Computer Science, University of Saarbrücken, 1990.

[Stock 91] 0. Stock. Natural Language and Exploration of an Information Space: the ALFresco Interactive System. Technical Report, Istituto per la Ricerca Scientifica e Tecnologica, Trento, Italy, 1991.

[Van Dijk 80] T. van Dijk. Textwissenschaft. München: dtv, 1980.

[Wahlster 91] W. Wahlster. User and Discourse Models for Multimodal Communication. In: J. Sullivan and S. Tyler (Eds.). Architectures for Intelligent User Interfaces: Elements and Prototypes. Reading, MA: Addison-Wesley, 1991.

[Wahlster&Kobsa 89] W. Wahlster and A. Kobsa. User Models in Dialog Systems. In: A. Kobsa and W. Wahlster (Eds.). User Models in Dialog Systems. Symbolic Computation Series, Berlin: Springer-Verlag, pp. 4-34, 1989.

[Wahlster et al. 89] W. Wahlster, E. André, M. Hecking, and T. Rist. WIP: Knowledge-based Presentation of Information. Report WIP-1, German Research Center for Artificial Intelligence, Saarbrücken, 1989.

[Wahlster et al. 91] W. Wahlster, E. André, W. Graf, and T. Rist. Designing Illustrated Texts: How Language Production Is Influenced by Graphics Generation. In: Proc. of the 5th Conference of the European Chapter of the ACL, Berlin: Springer-Verlag, 1991.