Intelligent Multi Media Presentation of information in a semi-immersive Command and Control environment

Cécile Paris, Nathalie Colineau
Commonwealth Scientific and Industrial Research Organisation (CSIRO)
ICT Centre
Sydney, NSW, Australia
Cecile.Paris@csiro.au
Nathalie.Colineau@csiro.au

Dominique Estival
Defence Science & Technology Organisation (DSTO)
Human Systems Integration Group
Command and Control Division
Edinburgh SA 5111, Australia
Dominique.Estival@dsto.defence.gov.au

Abstract

We describe the framework for an intelligent multimedia presentation system we designed to be part of the FOCAL laboratory, a semi-immersive environment for Command and Control Environment. FOCAL comprises a number of input devices and output media, animated virtual conversational characters, a spoken dialogue system, and sophisticated visual displays. These need to be coordinated to provide a useful and effective presentation to the user. In this paper, we describe the principles which underlie intelligent multimedia presentation (IMMP) systems and the design of such a system within the FOCAL multi-agent architecture.

1 Introduction

1.1 Description of FOCAL

FOCAL (Future Operations Centre Analysis Laboratory) was established at the Australian Defence Science and Technology Organisation (DSTO) to "*pioneer a paradigm shift in command environments through a superior use of capability and greater situation awareness*". The facility was designed to experiment with innovative technologies to support this goal, and it has now been running since 2000.

FOCAL contains a large-screen, semi-immersive virtual reality environment, where large quantities of information can be displayed. A number of modalities and media are available to display the information to the end-user. These include visual display mechanisms, such as 3-D virtual batttlespace, and spoken dialogue interaction with virtual conversational characters (VCCs) that allow presentation of information through speech as well as through textual displays (Taplin *et al.* 2001; Broughton *et al.*, 2002; Estival *et al.*, 2003). While these have so far been studied and implemented somewhat independently of each other, ultimately all the different available means to present the information to the end-user must work together and be combined into a coherent whole; otherwise, the result would be very confusing to the user.

From the delivery perspective (as opposed to the input fusion aspect) with which we are concerned here, FOCAL can be considered as an instance of an intelligent multimedia presentation (IMMP) system (see Bordegoni *et al.*, 1997 for a reference model).

1.2 An IMMP Architecture for FOCAL

The framework for the design of an intelligent multimedia presentation system (IMMP) within FOCAL was the result of a collaboration between DSTO and CSIRO. The aim was to design an architecture for the information delivery component, taking into account the existing architecture for the overall system, the available data sources and the type of desired presentations.

One of the main idea in FOCAL is that a VCC will serve as a Virtual Adviser (VA) to the team of commanding officers engaged in the planning or conduct of an operation. The aim of the VA is to engage in interactions with the officers, presenting information and offering advice. VAs are able to present the information and justify their advice through multimedia presentations (e.g., speech, video, text, map, etc.).

The remainder of this paper is structured as follows: in Section 2, we first briefly explain how research in multimedia presentation has grown from notions and systems developed in natural language generation. We then describe the process of generating multimedia

Proceedings of the Australasian Language Technology Workshop 2004
Macquarie University, Sydney, December 8th, 2004.

Accepted after full paper review

presentations, and, in particular, an approach to integrate coherently multimedia content, with examples from the FOCAL scenario. In Section 3 we describe the design for an IMMP architecture based on the reference model for FOCAL. We conclude in Section 4 with a short discussion of the evaluation to be undertaken.

2 IMMP Systems

Intelligent multimedia presentation systems (IMMP) are characterised by their capacity to automate the design of multimedia presentations. IMMP systems typically base their design decisions on explicit representations of diverse knowledge, and combine mechanisms and techniques that select, organise and coordinate relevant information across appropriate media. Such systems present the advantages to be:

- Adaptable and flexible by generating on-the-fly multimedia presentations of various combinations of information and media characteristics;
- Consistent by coordinating content within and across media, thus maintaining the coherence of the presentation; and,
- Effective by designing presentations that take into consideration the characteristics of the information source, the task that the users need to perform and the communicative goals to be achieved.

We first provide in Section 2.1 an overview of the principles that have guided the research in multimedia information presentation and describe in Section 2.2 the standard Reference Model for IMMP. We then present the process of generating multimedia presentations proposed by Colineau and Paris (2003), highlighting the main steps. In Section 2.3, we discuss the various issues encountered in integrating information across multiple media and illustrate the approach with an example from the FOCAL scenario.

2.1 Background

Studies in natural language generation have considerably influenced the research directions in multimedia information presentation, in particular on the issues of how to represent the global discourse structure, and how to organise and integrate each source of information in relation to the others. Several important notions have contributed to the progresses made in this domain:[1]

- the notion of discourse structure and the generation of multi-sentential texts, as embodied, for example in (McKeown, 1985a; 1985b; Moore and Paris, 1993);
- the notion of coherence and the rhetorical dependencies between discourse parts, as defined, for example, in Rhetorical Structure Theory (RST) (Mann and Thompson, 1988); and finally,
- the hierarchical planning approach as a means to structure and to represent a discourse goal hierarchy and the relationships between them, as in (Hovy, 1988; Moore and Paris, 1993) inter alia.

Starting from these notions, the generation of multimedia information presentations has been considered by many researchers (e.g., André and Rist, 1990; 1993; Maybury, 1993; Bateman *et al*., 1998; Green *et al*., 1998; Mittal *et a*l., 1998) as a goal-directed activity that starts from a communicative goal (i.e., a presentation intent), which is then further refined into communicative acts. Indeed, based on studies done in linguistics and philosophy (e.g., Austin 1962; Searle, 1969), in discourse (e.g., Grosz and Sidner, 1986) and in text planning (e.g., Hovy, 1988; Arens *et al*., 1993; Moore and Paris, 1993), the multimedia generation community has built on the idea that the internal organisation of a discourse or a presentation is composed of a hierarchy of communicative acts, each act supporting a specific communicative goal that contributes to the whole. It has then extended this principle to multimedia material. Thus, as pointed out by Maybury (1993, p.61):

"*As text can be viewed as consisting of a hierarchy of intentions, similarly, multimedia communication can be viewed as consisting of linguistic and graphical acts that, appropriately coordinated, can perform some communicative goal*".

Consequently, a question arises as to how to coordinate linguistic acts with other forms of expression (picture, graphics, video, etc.), so that the communicative goal is achieved in a coherent and consistent manner.

[1] See (Colineau and Paris, 2003) for details.

Accepted after full paper review

2.2 A Reference Architecture for IMMP

In recent years, a standard Reference Model (RM) for IMMP systems has been proposed by Bordegoni *et al*. (1997), aiming at providing a conceptual design of IMMP systems. The architecture is decomposed into five layers as follows:

§ The Control Layer controls the generation process by prioritising the communicative goals to be processed;

§ The Content Layer organises the content and makes explicit the relationships between discourse segments. It selects relevant information and chooses the appropriate modalities and media to be employed to convey the information and best achieve the communicative goals;

§ The Design Layer distributes to dedicated media/modality design modules communicative acts to be encoded. It also determines the spatial and temporal arrangements of media objects in the presentation. The design plan specifications produced for media objects are then passed onto the realisation layer;

§ The Realisation Layer distributes the design plan specifications to dedicated modules for the production of specific media objects. Specifications of displayable media objects with layout prescriptions are finally given to the presentation display layer; and,

§ The Presentation Display Layer combines media objects, defines the document or the display layout and finally delivers the multimedia presentation through specialised media devices. The result is a coordinated fusion of the output of the different devices. Here, Bordegoni *et al.* point out a clear distinction between the design and the production of media objects and their presentation.

We will not discuss here the Control Layer, as in the FOCAL system it is integrated with the overall dialogue and interaction management process (see Section 3). The other four layers are illustrated in Figure 1 and constitute the actual multimedia generation process.

The main steps that drive the multimedia presentation design are:

- *The content planning*: this stage aims at selecting and organising the content of the presentation. A discourse structure is produced, which makes explicit the role of each piece of content regarding to the whole presentation.
- *The media allocation and content realisation*: this stage aims at specifying how the content should be presented. One has to decide on the best way to realise the content and to combine the different discourse parts in a unified and integrated whole. We group here both the "Design of the presentation structure" and the "Realisation of the media objects"; and
- *The layout planning*: this stage aims at assigning location to content, grouping and aligning element of content to contribute to the legibility and readability of the presentation. For dynamic presentations, there is also a need to program the execution of the presentation, in particular setting the timing of all components.

In this paper, we focus on the content planning stage, but interested readers are referred to (Colineau and Paris, 2003) for details about the other stages.

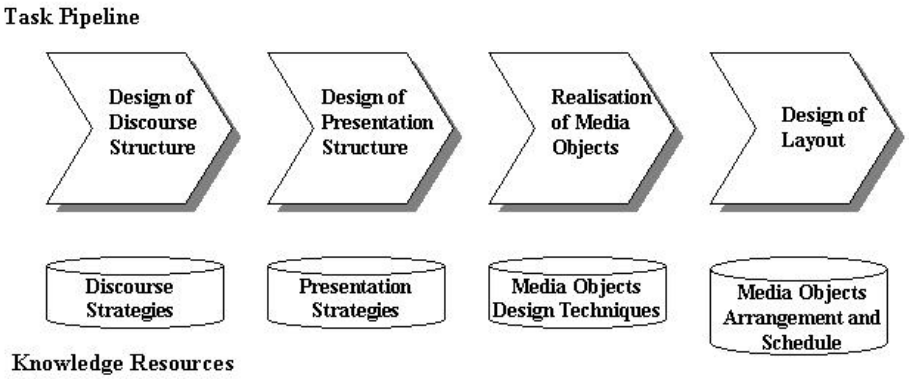

Figure 1: Multimedia generation process

Accepted after full paper review

2.3 Content Planning

When dealing with multimedia presentations, a number of issues that do not occur in simple text planning arise:

- How can we maintain the coherence of a presentation when the content is realised through different modalities (i.e., language, graphics, video, etc.)?
- How do graphical representations, animations, etc. work? Do they have an internal structure (as text does) that can be expressed in term of rhetorical and discursive dependencies?
- Can we use a common representation to express both textual and graphical acts?

Following research in the field of text generation, most multimedia information presentation systems have taken a unified approach, based on hierarchical planning, to structure and organise multimedia data. In parallel, by applying the principle of textual coherence to multimedia information presentation, researchers have generalised the RST theory of coherence to the broader context of multimedia information.

Using this theory, the organisation of the document or the presentation is represented by a tree structure (i.e., the document discourse structure). It is the output of the content planner, and it provides a detailed representation of the content to be produced, indicating how parts of the structure are related and which purposes different parts of the generated content serve (see Figure 2). In particular, this permits an explicit representation of the relationships and dependencies between discourse segments, whichever modalities and/or the media are selected afterwards.

The discourse structure[2] shown in Figure 2 illustrates the discourse representation that might be built to represent and organise the content of an "induction brief executive summary", from a military plannning exercise (with a fictitious scenario and fictitious data). This structure organises the different content elements (e.g., executive summary sentences, maps) and highlights their respective roles within the presentation (e.g., providing background information or evidences supporting a claim). We see that this executive summary is an integrated combination of text and illustrations (potentially static or dynamic illustrations). This example shows that the discourse structure may represent text as well as other multimedia contents, and that it can explicitly represent relationships across modalities (e.g., an illustration that supports text) and within modality (e.g., text that elaborates on another text part). If we examine the top of the discourse tree, it is organised into three discourse segments:

- the main node, which is a complex discourse segment considered as the nucleus (segment [2-5] + additional illustrations); and,
- two other discourse segments considered as satellites. One of the satellites (segment [1]) is linked to the nucleus by the rhetorical relation called *preparation*. This relation indicates that the satellite presents information which introduces the content presented by the nucleus. The other satellite is a complex discourse segment linked to the nucleus by the *elaboration* relation,[3] which indicates that the satellite provides additional information (e.g., geographic illustrations).

The discourse structure that is produced at the end of the content planning process presents several advantages. It provides a rich structure that can be reasoned about for a number of purposes, e.g., appropriate realisation in language, the placement of hypertext links, reasoning about user feedback and prior discourse, the coordination (as opposed to juxtaposition) of text, image, graphics or video.

Using a hierarchical planning approach ensures the unity of the whole multimedia information presentation by organising the entire presentation as one discourse structure, even though subparts may correspond to elements to be realised in different modalities and/or media. Having one overall discourse structure enables and facilitates the integration of various discourse elements.

[2] The discourse tree has been simplified for readability.

[3] Depending on the role of the satellite and the purpose of the information, the link between the satellite and the nucleus could also have been realised by the *enablement* relation. In that case, the information carried by the satellite would have supported the hearer in locating the region discussed in the summary.

Accepted after full paper review

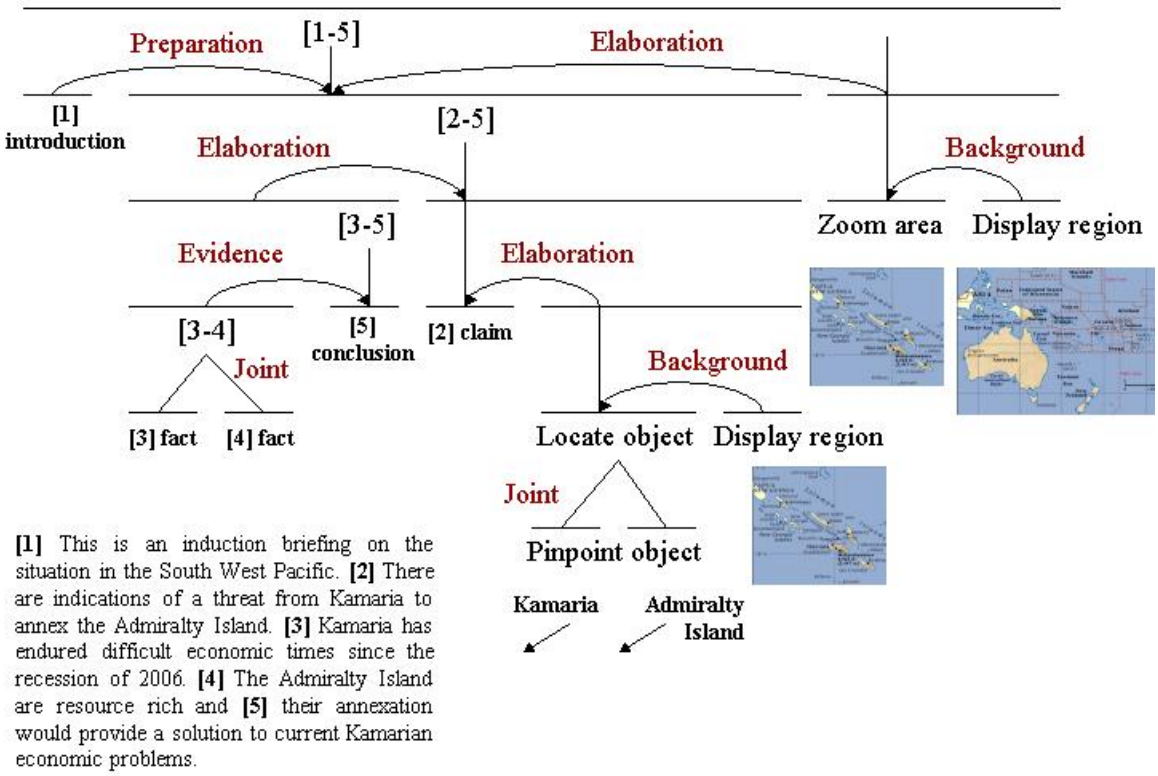

Figure 2: Example of multimedia content represented with RST (a fictitious scenario)

It also allows cross-references from one modality to the other (e.g., from text to graphics). This is explicitly stated in André and Rist (1995, p.9):

"*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.*"

This integrated representation enables the delivery component of the system to act as a media coordinator in the preparation of the final presentation script, ensuring for example that parts of the presentation are not duplicated. It then becomes possible to factor out the needs of each individual presentation segment and to share the media objects throughout the presentation. This integrated view of the multimedia presentation also ensures that parts of the presentation are coherent and well integrated with each other. With this approach, we can set and evaluate some basic multimedia principles, such as the principle of modality, contiguity, coherence and redundancy.

3 FOCAL: a Command and Control Environment

We now describe how we have extended the original FOCAL architecture to support the generation of multimedia presentations for military planning information.

3.1 Multimedia Presentation in the Focal Architecture

FOCAL is based on a multi-agent architecture, implemented using ATTITUDE, a high-level language developed at DSTO (Lambert and Relbe, 1998). ATTITUDE is capable of representing and reasoning with uncertainty about multiple alternative scenarios (Lambert, 1999). Extending the original FOCAL architecture (Taplin *et al.* 2001) for IMMP involved adding agents to explicitly handle the design, the composition and the realisation of multimedia objects. The original "Conductor" agent, which had so far only been concerned with spoken input, was renamed Dialogue Manager (DM). It is now responsible for dialogue flow control and for understanding users' query, deciding what best answers the user's needs (i.e., the communicative goal). A new MultiMedia Presenter (MMP) agent has been introduced. It organises the presentation of information within FOCAL and carries out most of the multimedia generation process shown in Figure 1. The third stage (i.e., the realisation of media objects) is left

Accepted after full paper review

to specific media generators, such as the natural language generator or the virtual video generator as shown in Figure 3. The data to be presented is accessed through the MMP agent, which determines, as part of the discourse plan, what information to include and integrate. The data come from various and heterogeneous sources, including spoken and typed input.

Figure 3 shows the FOCAL architecture from a presentation of information point of view, leaving aside aspects related to the processing and fusion of the input and connections to the Information Sources. It shows the different components and the main interactions amongst them. The architecture is organised around the two main agents for IMMP: the DM agent, responsible for the overall interaction, and the MMP agent, responsible for building a presentation and realising it using different media. They both act as "conductors": one for understanding, the other for generation.

Comparing the new architecture for FOCAL with the reference architecture for IMMP proposed by Bordegoni *et al.* (1997), the DM agent can be seen as corresponding to the Control Layer, deciding which communicative goals should be processed (i.e., the purpose of the presentation), while the MMP agent assumes the processes performed by the Content and Design layers.[4]

3.2 Interaction flow process

In the current FOCAL scenarios, there are two modes: (1) the virtual adviser (VA) "pushes" the information that needs to be presented, namely delivers the briefing content, and (2) the VA allows users to ask questions to repeat or gain information.

With the IMMP architecture for FOCAL shown in Figure 3, these two modes follow the same flow process. In both cases, the aim is to answer either an explicit or an implicit information need by presenting information through complementary media. The information need may have been initiated by the system (i.e., briefing mode) or initiated by the user (i.e., question-answering/dialogue mode).

When the system is answering a user's query, the DM agent has to understand the user's query in order to identify what is the user's information need.[5] This requires the DM agent to have access to domain knowledge, e.g., an ontology of the domain (see Nowak *et al.*, 2004), to ensure that the query makes sense (i.e., is syntactically and semantically well-structured). Then, the DM's aim is to determine a communicative goal which answers this information need and to send this goal to the MMP agent. The communicative goal thus constitutes the input to the MMP agent. From this input, the MMP selects the appropriate discourse strategies to be developed (e.g., "explain mission").

Once the MMP receives a communicative goal, it can develop a discourse plan to satisfy this goal. The discourse plan aims at selecting the relevant content and organising it. A set of queries is thus sent to the Query agent to acquire the content identified.

Depending of the level of knowledge and expertise of the Query agent, the queries can either be forwarded to a specific Information Source (IS) agent responsible for the information requested, or be forwarded to all IS agents. In the latter case, the Query agent will have to choose the most appropriate amongst the responses received and send these to the MMP. The Query agent thus acts as an interface between the IS agent and the MMP agent. When the content of the information to be presented has been retrieved, the MMP allocates the realisation of each discourse segment (i.e., presentation unit) to the media-specific generators. The decision to encode information under a particular modality is made by taking into account several criteria represented as declarative rules and used by the planner engine.

Finally, the MMP has to supervise the realisation of each discourse segment. It acts as a media coordinator to ensure that each media-specific generator agent is working towards a consistent and synchronised presentation. The MMP ensures that the presentation plan is built cooperatively and that alternatives are negotiated if needed. Thus, the presentation design planning is a cooperative process amongst the media-specific generator agents, supervised by the MMP. In comparison with the reference architecture model, the MMP shares with the media-specific generator agents the tasks performed in the design layer of the architecture.

Each media-specific generator agent receives a discourse segment to be realised. It develops the design of this segment closely with the MMP before starting the generation process. These agents use specific knowledge sources (e.g., grammar and lexicon, icons region models, texture models, graphics techniques, etc.). In comparison with the reference architecture model, the media-specific generator agents perform the tasks represented in the realisation layer of the architecture.

[4] We are currently collaborating with UniSA on the design of a Media Selection agent and a Media Presentation agent for these two layers.

[5] In briefing mode, the DM agent generates the information need, while in dialogue mode, the information need comes from the input devices, whose output is sent to the Input Fuser and then to the DM.

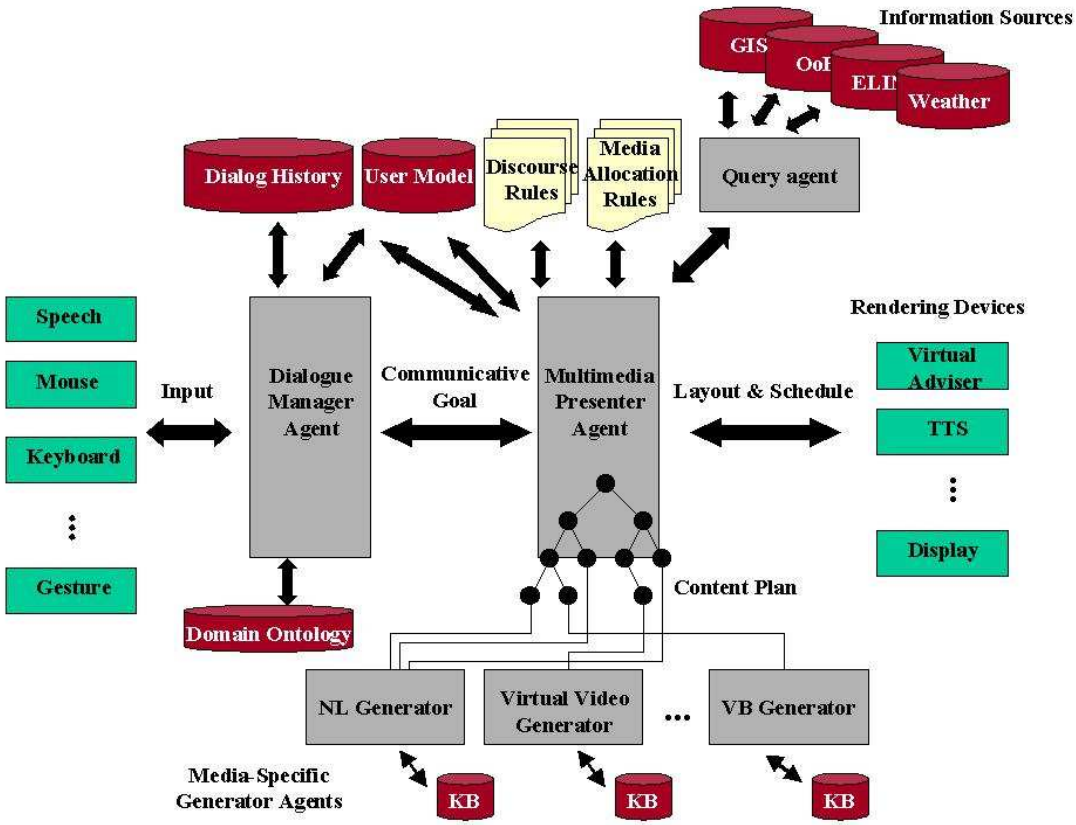

Figure 3: IMMP architecture for FOCAL

When each presentation unit has been realised and appropriately scheduled on a single timeline by the MMP, they are sent to their specific rendering devices to be displayed (cf. the presentation layer of the reference architecture).

The architecture and the interaction process flow described above have been designed to handle an interaction between a user and the FOCAL system. This means that, during a session, a user may interact with several virtual advisers. Virtual advisers can be considered as a means to interact with the system in the same way as a mouse or a pointing device. In this case, one DM and one MMP drive the interaction and provide appropriate answers. However in the case of multiple users interacting simultaneously with the system and in particular with different virtual advisers, the architecture will need to be extended to support parallel interactions. It will be necessary to have one DM and one MMP per interaction stream.

4 Discussion

This work is still in its early stages, and the architecture proposed here has not yet been fully implemented; however the current FOCAL system is very much in line with the IMMP architecture. Although attention has so far been put mainly on the spoken dialogue with the virtual advisers (Estival *et al.*, 2003) and on the integration of new input modalities within a unified framework (Wark *et al.*, 2004), our intention is to continue the work to produce appropriately integrated presentations. To conclude this paper, we would like to briefly discuss issues of evaluation.

The FOCAL environment may be evaluated at different levels and for different purposes. The aspect which concerns us here is the generation of multimedia information and its integration across several modalities or media. The important questions then are whether users receive enough information, whether the information is relevant for them to accomplish their task, and whether the information has been appropriately represented and integrated. Through evaluation, we would like to be able to answer questions such as:

- Is a particular medium/modality to be preferred for the encoding of specific information in order to facilitate the comprehension and retaining of that material? Which information should be represented under which format?
- Which modalities best complement each other?
- How can we avoid the split-attention effect in multimedia material?
- Does the verbal or visual representation of information have an impact on its processing by users?

Accepted after full paper review

5 Acknowledgements

We would like to thank our colleagues in the FOCAL team, Dr Steven Wark, Michael Broughton and Andrew Zschorn for their invaluable contribution to this project. We also wish to acknowledge the support of Nuance for the development of the speech recognition system.

References

André, E. and Rist, T. (1990) Towards a plan-based synthesis of illustrated documents. In *Proc. of 9th ECAI*, Stockholm, Sweden, 25-30.

André, E. and Rist, T. (1993) The design of illustrated documents as a planning task. In M. Marbury (Ed.), *Intelligent Multimedia Interfaces*, ch 4, 94-116. AAAI Press / The MIT Press.

Arens, Y., Hovy, E. and van Mulken, S. (1993). Structure and rules in automated multimedia presentation planning. In *Proc. of the 13th International Joint Conference on Artificial Intelligence (IJCAI'93)*, Chambery, France.

Austin, J. (1962) *How to do things with words*. Ed. J.O. Urmson. England: Oxford University Press.

Bateman, J., Kamps, T., Kleinz, J. and Reichenberger, K. (1998). Communicative goal-driven NL generation and data-driven graphics generation: an architectural synthesis for multimedia page generation. In *Proc. of the 9th International Workshop on Natural Language Generation, Niagara-on-the-Lake, Canada*.

Bordegoni, M., Faconti, G., Maybury, M.T., Rist, T., Ruggieri, S., Trahanias, P. and Wilson, M. (1997). A standard reference model for intelligent multimedia presentation systems. In *Computer Standards and Interfaces: the International Journal on the Development and Application of standards for computers, Data Communications and Interfaces*, 18(6-7).

Broughton, M., O. Carr, D. Estival, P. Taplin, S. Wark and D.Lambert (2002). "Conversing with Franco, FOCAL's Virtual Adviser". *Human Factors 2002*, Melbourne.

Colineau, N. and Paris, C. (2003) Framework for the Design of Intelligent Multimedia Presentation Systems: *An architecture proposal for FOCAL*. CMIS Technical Report 03/92, CSIRO, May 2003.

Estival, D. Broughton, M., Zschorn, A. and Pronger, E. (2003). Spoken Dialogue for Virtual Advisers in a semi-immersive Command and Control environment. In 5th *SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan, pp 125-134.

Green, N., Carenini, G., Kerpedjiev, S., Roth, S. and Moore, J. (1998). A media-independent content language for integrated text and graphics generation. In *Proc. of the COLING'98/ACL'98 Workshop on Content Visualization and Intermedia Representations* Montréal, Canada.

Grosz, B. and Sidner, C. (1986). Attention, Intentions, and the structure of discourse. In *Computational Linguistics* 12(3), 175-204.

Hovy, E. (1988). Planning coherent multisentential text. In *Proc. of the 26th Conference of the ACL*, Buffalo, NY, 163-169.

Lambert, D. and Relbe, M. (1998). Reasoning with Tolerance. In *Proc. of the 2nd International Conference on Knowledge-Based Intelligent Electronic Systems*. IEEE. pp. 418-427.

Lambert, D. (1999). Advisers With Attitude for Situation Awareness. In *Proc. of the 1999 Workshop on Defence Applications of Signal Processing*. pp.113-118, LaSalle, Illinois.

Mann, W. and Thompson S. (1988). Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text*, 8(3), 243-281.

Maybury, M. (1993). Planning multimedia explanations using communicative acts. In M. Marbury (Ed.), *Intelligent Multimedia Interfaces*, chapter 2, 59-74. AAAI Press / The MIT Press.

McKeown, K. R. (1985a). Discourse strategies for generating natural-language text. *Artificial Intelligence*, 27(1):1-42.

McKeown, K.R. (1985b). *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press, Cambridge, England.

Mittal, V., Moore, J., Carenini, G. and Roth, S. (1998). Describing complex charts in natural language: a caption generation system. In *Computational Linguistics, Special issue on Natural Language Generation. Vol. 24, issue 3*, 431-467.

Moore, J. and Paris, C. (1993). Planning text for advisory dialogues: Capturing intentional and rhetorical information. In *Computational Linguistics*, vol.19, Number 4, 651-694.

Nowak, C, D. Estival and A. Zschorn (2004). "Towards Ontology-based Natural Language Processing". RDF/RDFS and OWL in Language Technology: 4th Workshop on NLP and XML (NLPXML-2004), ACL 2004, Barcelona, Spain. pp-59-66.

Taplin, P., G. Fox, M. Coleman, S. Wark and D. Lambert (2001). "Situation Awareness Using A Virtual Adviser". OzCHI.

Wark, S., A. Zschorn, M. Broughton and D. Lambert. (2004). "FOCAL: A Collaborative Multimodal Multimedia Display Environment". *SimTecT 2004*, Canberra, Australia. pp-298-303.