MULTIMODAL AND MULTIMEDIA SYSTEMS: ARCHITECTURES FOR ADVANCED DIALOGUE.

M.D. Wilson,
SERC Rutherford Appleton Laboratory, Oxon., UK.

P. Falzon
INRIA, Roquencourt, France

ABSTRACT

Multimodal systems use a single meaning representation language for all information and choose the effective way to present this for a specific user at a point in task performance through the available media. In contrast, many multi-media systems retrieve information stored in media specific representations which constrain the presentation options. This paper investigates the trade-off of storage and retrieval efficiency against presentation effectiveness for advanced dialogue systems.

INTRODUCTION

Recently several demonstrators of multimodal systems (see Wilson and Conway, 1991) have been produced. These make a strong commitment to “multimodal” rather than to “multimedia” interaction in the interface. The distinction intended is that a multi-media system is one which uses different presentation media (e.g. text, raster graphics, video, speech) without a commitment to the underlying representation of the information presented. For reasons of efficiency of both storage and processing, individual specialised representations are usually used for information which is intended to be presented in each individual medium, and information can only be presented in a single medium. A “multimodal” system is one which includes several input and output media, but is committed to a single internal representation language for the information. This permits the same information to be presented in any mode, chosen purely by rules which select that mode for that user at that point in task performance as being both sufficiently expressive and most efficient.

The MMI^2 system (Man-Machine Interface for MultiModal Interaction with knowledge based systems) will be used as an exemplar of a multimodal system (Binot et al., 1990; Ben Amara et al., 1991; Wilson et al., 1991). The modes available in the MMI^2 demonstrator are: for input: English, French and Spanish natural languages, gesture, direct manipulation of graphics, command language; and for output: English, French and Spanish natural languages, graphics (CAD diagrams and business graphics), and non-verbal audio. The meaning representation language used for all information within the system is called the Common Meaning Representation (CMR). This includes is a typed first order logic with relativised quantification and second order relation symbols as well as the promiscuous reification of objects and events (after Hobbs, 1985).

The architecture of the MMI^2 system can be described as the three layers of Seehiem model for UIMS design (Pfaff, 1985; Duce et al., 1991). The top layer contains the input and presentation modes, the middle layer is the dialogue management layer, and the bottom layer is the application knowledge based system (see Figure 1).

If we consider an application about computer networks, a user may ask the natural language question (1). The application would provide the information described in (2) associating a machine with a cost in pounds Sterling (the CMR representations of these statements are given in Appendix 1). The communication planning expert would determine that both natural

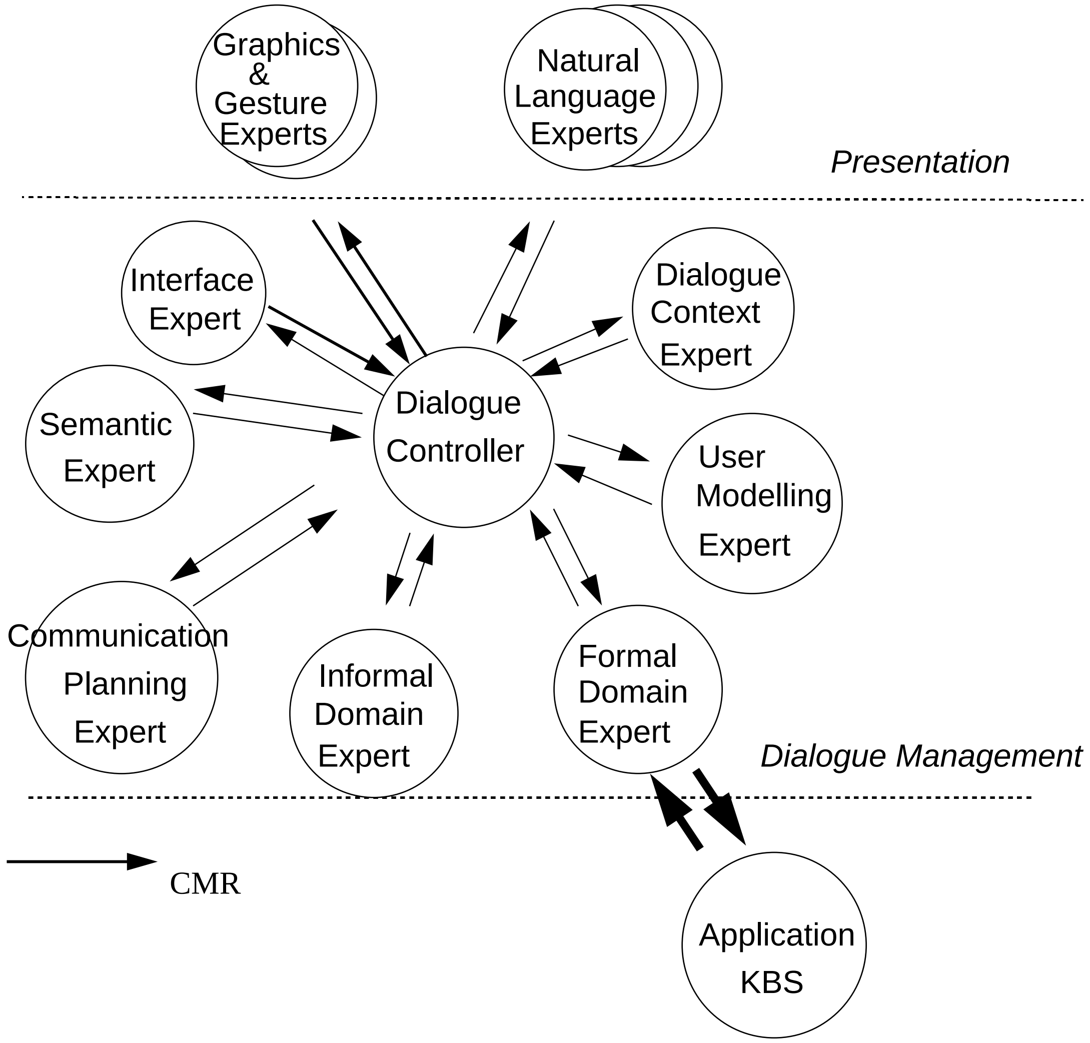


Figure 1: Architecture for the first MMI^2 demonstrator

language and graphics modes could express this information, but that graphics mode will more efficiently present the information for the user to understand (by a simple heuristic that sequences of paired associations are better understood in graphical form). The Graphics Expert would then choose which presentation tool to use for this information between pie charts, bar charts, line graphs, hierarchies etc.. It would draw on information about the user (from the User Modelling Expert), the task (from the Domain Expert), and the context (from the Dialogue Context Expert) in order to make this decision (after Mackinlay, 1986). It would then design the presentation object (chart, table etc.) and pass the required information in the data structure required by the appropriate presentation tool.

(1) What do the computers on the network cost?

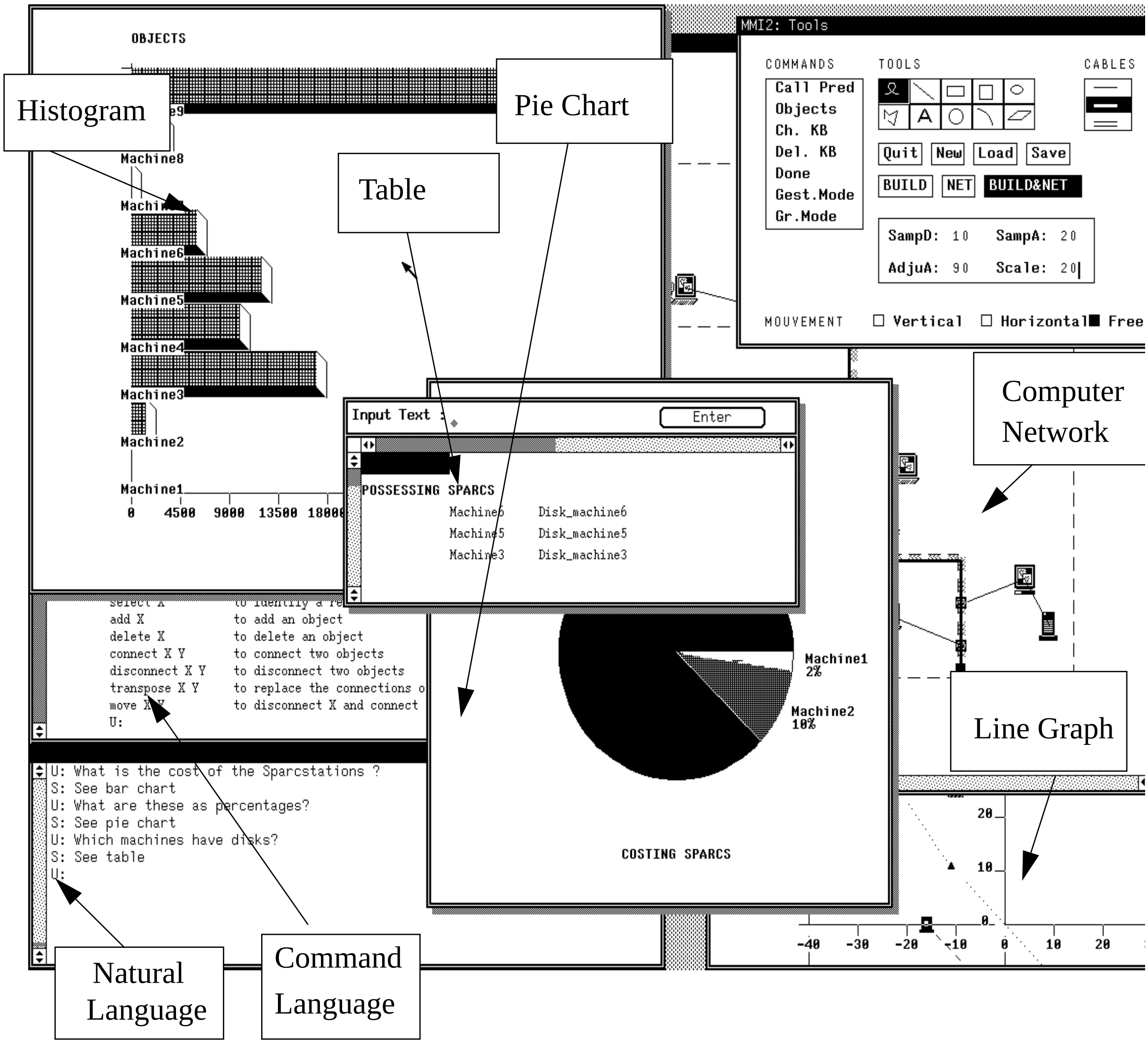
(2) Machine1 costs 5113; Machine2 costs 9208; Machine3 costs 5113; Machine4 costs 30625; Machine5 costs 9208; Machine6 costs 30625; Machine7 costs 43750; Machine8 costs 5113.

This simple example is possible in many information retrieval systems; examples where answers are presented in a combination of natural language and changes to a CAD diagram of a building and network show the individual power of MMI^2, but they require more space than is available here to describe. The two principles behind mode integration in this system are:

A) mode integration should mainly be achieved by an integrated management of a single generalised discourse context.

B) there is a meaning representation formalism, common to all modes, which is used as a vehicle for internal communication of the semantic content of interactions inside the interface and also used as a support for semantic and pragmatic reasoning.

.



```
select X            to identify a re
add X               to add an object
delete X            to delete an object
connect X Y         to connect two objects
disconnect X Y      to disconnect two objects
transpose X Y       to replace the connections o
move X Y            to disconnect X and connect
U:
```

```
U: What is the cost of the Sparcstations ?
S: See bar chart
U: What are these as percentages?
S: See pie chart
U: Which machines have disks?
S: See table
U:
```

```
POSSESSING SPARCS
          Machine6     Disk_machine6
          Machine5     Disk_machine5
          Machine3     Disk_machine3
```

SampD: 10   SampA: 20
AdjuA: 90   Scale: 20

MOUVEMENT   □ Vertical   □ Horizontal   ■ Free

Figure 2: A typical screen display of MMI^{2}in use.

Card et al., (1991) have presented an analysis of information retrieval tasks in terms of the cost ratios. The cost of information can be considered as the sum of the cost of retrieving the information and the cost of assimilating it. This is reflected in the use of desks, local shelves and distant files in the office environment. Multi-modal systems such as MMI^2 attempt to reduce the cost of assimilation of information by aiding query formation, and selecting effective and efficient presentation formats by considering information access as part of a larger task context. However, they do not address either the cost of retrieval or the cost of storage. Multimedia database systems aid assimilation in as far as the information is stored in a predetermined medium which is seen to well suited to its use; but they also address retrieval costs by using specialised retrieval mechanisms to for each medium.

An improvement over both situations would be to reduce retrieval costs through the use of specialised media storage and retrieval whilst also allowing the tuning of these media to the user in a task context within the parameters available for each medium. It is therefore necessary to use establish a single generalised discourse context where the denotation of context symbols are objects in individual media, and provide a language which will permit reasoning over both the content of these symbols and the media constraints upon them. This requires a typing system in the meaning representation formalism which applies to both content and to form.

Developments in SGML and HY-TIME as standards for describing textual objects but also objects in other media (e.g. audio, motion video, static images) are an attempt to provide such a type system. The rules for relating tag types define the intensional definitions appropriate for the content, while the rules for presenting objects of tagged types can be dynamically modified on the basis of task, user and dialogue context information to improve assimilation.

Within these constraints it is then necessary to reformulate the presentation construction rules used in multimodal systems so that they are bound not only by the task, dialogue context and user, but also by the medium of the information to be presented. What is required is a set of rules of the sort used in multimodal systems for all information which are constrained to individual media, but still express the full range of options for presentation to reduce the cost of assimilation by the user.

REFERENCES

Ben Amara, H., Peroche, B., Chappel, H., Wilson, M.D. (in press) Graphical Interaction in a Multi-Modal Interface. In Proceedings of Esprit ‘91 Conference, Kluwer Academic Publishers: Dordrecht.

Binot, J-L., Falzon, P., Perez, R., Peroche, B., Sheehy, N., Rouault, J. and Wilson, M.D. (1990). Architecture of a multimodal dialogue interface for knowledge-based systems. In Proceedings of Esprit ‘90 Conference, 412-433. Kluwer Academic Publishers: Dordrecht.

Card, S.K., Robertson, G..G. and Mackinlay, J.D (1991) The information Visualizer, an information workspace, in Proceedings of ACM CHI ‘91 Conference, 181-188.

Duce, D.A., Gomes, M.R., Hopgood, F.R.A. and Lee, J.R. (1991) User Interface Management and Design, Eurographic Seminar Series, Springer-Verlag: Berlin.

Hobbs, J.R. (1985) Ontological Promiscuity. In Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics, Chicago, June 1985, 61-69.

Mackinlay, J. (1986) Automating the Design of Graphical Presentations of Relational Information, ACM Transactions on Graphics, 5(2), 110-141.

Pfaff, G.E, (1985) User Interface Management Systems, Eurographic Seminar Series, Springer-Verlag: Berlin.

Wilson, M.D. and Conway, A. (1991) Enhanced Interaction Styles for User Interfaces, IEEE Computer Graphics and Applications, vol 11, pp 79-90.

Wilson, M.D., D. Sedlock, J-L. Binot, P. Falzon, (1991) An Architecture For Multimodal Dialogue, in M.M. Taylor, F. Neel & D.G. Bouwhuis (Eds.),Proceedings of the second Vencona Workshop on Multi-Modal Dialogue, 1991.

APPENDIX 1

CMR representations of the example query and answer (1) and (2) used in the text.

(1) User input: What do the computers on the network cost?

```
CMR(
[
 CMR_act_analysis(
        u_type(wh([x1]),question_mark),
        [
         CMR_exp(
                [
                 anno(x2,[singular,definite,neuter]),
                 anno(x3,[plural,definite,neuter]),
                 anno(x1,[indefinite,singular])],
                description(desc(E,x2,LAN,true),
                description(desc(E,x3,COMPUTER,
                        description(desc(E,x5,IS_ON,true),
                        conj(
                                [
                                atom(PRESENT,[var(x5)]),
                                atom(ARG1,[var(x5),var(x3)]),
                                atom(ARG2,[var(x5),var(x2)])])))),
                description(desc(null,x1,COST,true),
                description(desc(E,x4,COSTING,true),
                conj(
                        [
                        atom(PRESENT,[var(x4)]),
                        atom(ARG2,[var(x4),var(x1)]),
                        atom(ARG1,[var(x4),var(x3)])]))))),
                nil)],
        nil)]
ok,
English,
time(51,42,20,11,07,1991))
```

(2) System response:

```
CMR([
        CMR_act_analysis(
                u_type(declarative,none),
                [
                CMR_exp(
                        [],
                        conj([           [
atom(COSTING,[const(Machine1),struc(COST,[const(5113),const(STERLING)])]),
atom(COSTING,[const(Machine2),struc(COST,[const(9208),const(STERLING)])]),
atom(COSTING,[const(Machine3),struc(COST,[const(5113),const(STERLING)])]),
atom(COSTING,[const(Machine4),struc(COST,[const(30625),const(STERLING)])]),
atom(COSTING,[const(Machine5),struc(COST,[const(9208),const(STERLING)])]),
atom(COSTING,[const(Machine6),struc(COST,[const(30625),const(STERLING)])]),
atom(COSTING,[const(Machine7),struc(COST,[const(43750),const(STERLING)])]),
atom(COSTING,[const(Machine8),struc(COST,[const(5113),const(STERLING)])])]),
                         nil)],
                nil)]
ok,
Graphics,
time(41,46,11,21,5,1991))
```