A Conversational Model of Multimodal Interaction in Information Systems

Adelheit Stein, Ulrich Thiel

German National Research Center for Computer Science
Integrated Publication and Information Systems Institute (GMD-IPSI)
Dolivostrasse 15, 6100 Darmstadt, Germany
Email: {stein, thiel}@darmstadt.gmd.de

Abstract

We propose a comprehensive framework for modeling and specifying multimodal interactions. To this end, we employ an extended notion of 'dialogue acts' which can be realized by linguistic and non-linguistic means. First, a set of constraints is presented that describes the temporal structure and all patterns of exchange during a cooperative information-seeking dialogue. Second, we introduce a strategic level of description which allows the specification of the topical structure according to an information-seeking strategy. The model was used to design and implement the MERIT system, and led to a reduction in the complexity of the user interface while preserving most of the useful, but sometimes confusing, dialogue options of advanced direct manipulation interfaces.

Introduction

Multimodal user interfaces of contemporary information systems use various means of conveying information, e.g. text, graphics, forms, tables, and pictures. If the presented information items become complex, in most systems users are allowed to investigate the items directly. Thus, the distance between the users' intentions and the objects is reduced to a minimum (cf. Hutchins, Hollan & Norman 1986).

However, this exploratory approach to information retrieval has to overcome the problems arising from browsing in a large information space. These difficulties, which are well known in hypertext applications, can be attacked by combining the direct manipulation of objects by the user with cooperative system responses. To date, cooperation has mostly been investigated in the context of natural language interfaces which regard user-machine interaction as a dialogue between two partners. As this notion is often referred to as the *conversation metaphor* (cf. Reichman 1986, 1989), the proposed hybrid interaction style, which integrates graphical and natural language components in a multimedial environment, will be called *multimodal conversation* in the following.

In this paper, we will introduce a comprehensive model of multimodal conversations. The next section outlines the notion of multimodal dialogue acts, whereas in the third section the constraints that govern the structure of a cooperative multimodal dialogue are discussed. In the fourth section, we present an overview of the prototypical information system MERIT, which was designed and implemented as an application of the conversational model. Some of the benefits of this approach are sketched in the concluding part of the paper.

The Notion of Multimodal Dialogue Acts

In the conversational approach, user inputs such as mouse clicks, menu selections, etc. are not interpreted as invocations of methods that are executed independent of the dialogue context. Instead, the direct manipulation of an object is considered to be a *dialogue act* expressing a discourse goal of the user. Therefore, the system can respond in a more flexible way by taking into account the illocutionary and semantic aspects of the user's input. Additionally, this approach to human-computer interaction provides a basis for the integration of different interaction styles, such as natural language and graphics, in a multimodal information system (cf. e.g. Arens & Hovy 1990, Feiner & McKeown 1990, Maybury 1991, Oei et al. 1992).

Based on a representation of the dialogue history, *dialogue acts* of the user, which are performed by directly manipulating the display structure, can be identified. For this reason, the user's manipulations (e.g. mouse clicks) are not only executed as methods sent to the graphical surface objects, but they affect transformations of the underlying internal representation of the ongoing dialogue.

The reactions of the system are instances of generic dialogue acts like 'inform', 'offer', or 'reject', which are modeled as frames. Since we use an object-oriented presentation style, a system's dialogue act is performed by creating or changing *informational objects*. An informational object represents a fragment extracted from the underlying database together with a specification of its graphical and/or textual presentation. In general, the system responds by visualization of new graphical objects on the screen in combination with the generation and presentation of a 'comment' in natural language.

Structure of Multimodal Dialogues

While the notion of dialogue acts captures the essence of single contributions to a multimodal interaction, a conversational model has also to address the problem of combining these actions to coherent sequences. Our approach takes local as well as global patterns of dialogues into account. The local structures are governed by the interrelations of *dialogue roles* and *tactics* of the information seeker and the information provider, and described by a complex network of interrelated generic dialogue acts. The global aspect relates the sequence of dialogue contributions to *information-seeking strategies*.

Thus, a system is able to plan the content of subsequent dialogue steps according to principles of topical coherency.

The Conversational Roles Network

A formal schema of information-seeking dialogues allows dialogues to be modeled at the discourse act level or – in terms of Speech Act Theory – the illocutionary level (cf. Searle & Vanderveken 1985).

In order to capture the temporal structure of the dialogue, we introduce the notion of *dialogue states*. During a dialogue, the dialogue acts performed by the dialogue partners change the dialogue state. Since there may be several possible continuations, our model for possible dialogues resembles a network, whereas each actual dialogue is represented as a sequence of singular acts. The network we developed for our problem domain of information-seeking dialogues is called COR (modeling "Conversational Roles"). For a detailed description of the formalism and the theoretical framework we refer to Sitter & Stein (1992), and Maier & Sitter (1992).

COR can be regarded as a recursive state-transition network like the "Conversation for Action Model" of Winograd and Flores (1986). In addition, our model for information-seeking processes adopts some concepts from Systemic Linguistic approaches to discourse modeling (cf. Fawcett et al. 1988, Halliday 1984) and Rhetorical Structure Theory (Mann & Thompson 1987). Basically, the COR network defines the generic dialogue acts available (e.g. asking, offering, promising, answering, evaluating), their possible interrelations, and the mutual role-changes of speaker and addressee.

Figure 1 shows the basic schema of COR: the circles represent *states* on the top-level of the dialogue, the squares terminal states. Arrows represent transitions between two states, i.e. the *dialogue contributions*. Parameter A refers to the information seeker, B to the information provider. The order of the parameters indicates the speaker - hearer roles; the first parameter indicates the speaker, the second the addressee.

Note, *first*, that the dialogue contributions (transitions) are themselves transition networks which may contain sub-dialogues of the type of the basic schema. We distinguish between two types of networks of dialogue contributions (cf. fig. 2 and fig. 3) which are described below. *Second*, one should keep in mind that in COR all dialogue contributions – except the inform contribution – can be 'implicit', i.e. they may be omitted (a 'jump') when the implicit intention can be inferred from the current context.

The bold arrows between the states <1> and <5> represent two 'idealized' straightforward courses of a dialogue:

- A utters a request for information, B promises to look it up (possibly skips the promise) and presents the information, A is satisfied and finishes the dialogue.
- B offers to provide some information (anticipating an information need of A), A accepts the offer (or part of it), B provides the information, A finishes the dialogue.

However, such simple courses of actions are very rare in more problematic situations. Participants often depart from such a straight course, and perceive their departure as quite natural. Information-seeking dialogues are also highly structured and normally contain a lot of corrections and clarification sequences. The interactions of the participants build a complex net of mutually related commitments and "role expectations" (cf. e.g. Halliday 1984). Simple question-answer dialogue models (like those applied in most of the classical interfaces to information systems) cannot cover this complexity.

Instead, we need a description of a flexible interaction which allows both dialogue partners to correct, withdraw, or confirm their intentions, and to insert clarifying sub-dialogues. To this end, we invented several transitions for with-

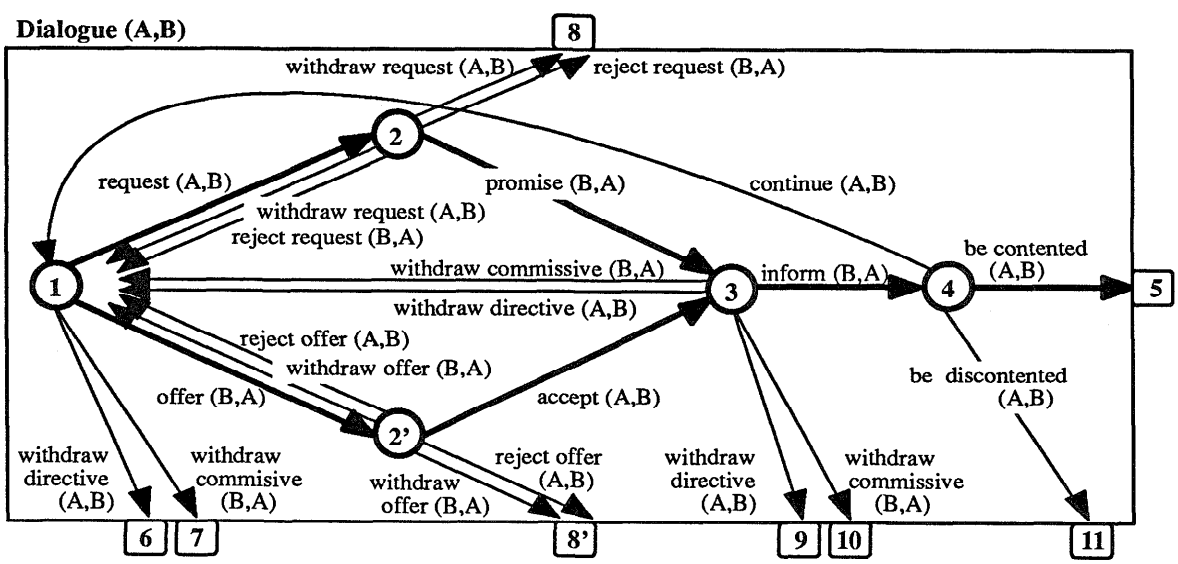

Figure 1: The basic COR 'dialogue' schema

directives: request, accept; *commissives:* offer, promise

drawing or rejecting a contribution. In every dialogue state <from 2 to 4>, A and B may return either to state <1>, thus preparing a new dialogue *cycle*, or they may quit the current dialogue (states <5-11>).

The embedding of clarifying sub-dialogues is described by the two networks below:

Figure 2 displays the schema for the 'inform' contribution, i.e. the transition between <3> and <4> in the basic dialogue network. A more general term is 'assert', indicating an assertion or a statement, but in our context of information-seeking dialogues we use the more specific term.

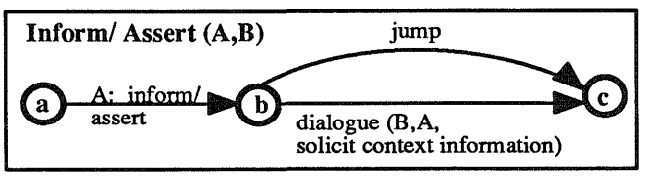

Figure 2: Schema of an 'inform' ('assert') contribution

A starts with an atomic inform (atomic acts are expressed by the notation A: ....). This inform act (its locution) could be quite a long monologue, i.e. a text, or a graphical presentation comprising several propositions. Of course, in that case it would have a semantic sub-structure and consist of several elements, such as sentences. But the illocutionary point would not change, i.e. no new commitments or expectations are expressed or imposed on the addressee. The atomic inform act can either lead directly to state <c> (jump to <c> and at the same time in the dialogue network to <4>). Or B may decide to initiate a sub-dialogue to solicit more context information about A's inform act, e.g. asking a question related to the inform act. This transition between <b> and <c> is a traversal of a basic dialogue network.

The network in figure 3 is more complex. All dialogue contributions, except 'inform', follow this pattern. When A intends, for instance, to make a request, she can follow one of the two possible paths: <a-b-c> or <a-b'-c>. On the first path A formulates the request and either 'jumps' to <c>, or appends an 'assert' (network of fig. 2), supplying voluntarily some context information related to the request. If this context information is not given by A, B might decide to start a sub-dialogue to solicit such context information, e.g. asking for details, or the background of the request. The other path is similar, but with a revised order. After an assert of A supplying some context information about the intended request (or jump), she can either formulate the request now explicitly, but also skip it (jump). The latter would create the situation of an indirect request which is reminiscent of the term "indirect speech act" coined in Speech Act Theory (cf. Searle 1975). A simple example is: A: *"I don't know much about this RACE funding program."* Even if A does not then ask directly *"What is it about?"*, B might infer that A has expressed an indirect request by the first statement.

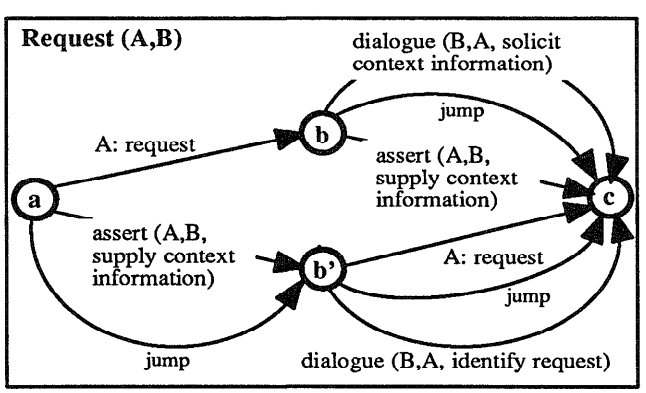

Figure 3: Schema of a 'request' contribution

We distinguish between two main types of components within these dialogue contribution networks (cf. in detail Sitter & Stein 1992). The first expresses the function (illocutionary point) of the whole dialogue contribution. This is normally an atomic dialogue act, e.g. the request or inform act in our example. The second serves for exchanging additional context information which is either supplied voluntarily (assert contributions) or requested in a sub-dialogue.

Using the terminology of Rhetorical Structure Theory – RST (cf. Mann & Thompson 1987) we call a component of the first type a "nucleus" and a component of the second type a "satellite". Both, nucleus and satellite, are related to one another by rhetorical or semantic relations. The set of relations described by RST was developed for the analysis of written texts, i.e. monologues, and has been extended in other approaches in the field of Computational Linguistics (overview in: Maier & Hovy 1993). However, there exist some recent attempts to combine research done in the text-linguistic field with dialogue modeling approaches (cf. Maier & Sitter 1992, Fawcett & Davis 1992). Maier and Sitter, for instance, extended the set of necessary relations, especially "interpersonal relations", for the dialogue situation and combined it with the COR approach. This proved to be very useful in our context of human-computer retrieval dialogues, because their specifications can be directly integrated in our application, the MERIT system.

Modeling Information-Seeking Plans

The COR network covers the illocutionary structure of dialogues, but does not supply means for a specification on the thematic level. Since the thematic level governs the selection of the contents communicated in the dialogue acts, it plays an essential role in dialogue planning. If we want the system to engage in a meaningful cooperative interaction with the user, we have to address this question by supplying a prescriptive addition to the – thus far descriptive – dialogue model. We perceive actual dialogues as instantiations of more abstract entities, each of which represents a class of concrete dialogues. This is similar to approaches to the generation of multimodal utterances using plan operators (e.g. Maybury 1991, Rist & André 1992). However, the abstract representations of dialogue classes are more complex. The classes comprise dialogues that show a similar basic pattern. Usually these patterns are closely related to certain strategies which are pursued by the user during her interaction. In the case of information systems, Belkin, Marchetti, and Cool

(1993) suggested a classification of dialogues with respect to information-seeking strategies. Based on this approach, we developed a set of typical dialogue plans or "scripts" (cf. Schank & Abelson 1977) for a given domain and task (cf. Belkin, Cool, Stein & Thiel 1993). A collection of dialogue plans is the basis for selecting an appropriate plan for a given information need. Once a plan has been chosen, it provides suggestions to the user on how to continue in a given dialogue situation, and specifies cooperative system reactions.

On the implementation level, we represent dialogues as sequences of *dialogue steps*. The internal structure of a dialogue step is given by two parameters: the *perspective* of the step, and its *implementation*. The perspective determines the topical spectrum that can be addressed in this step without destroying the thematical coherence of the dialogue in general. Similar notions have been proposed by McCoy (1986) in the area of natural language interfaces and Reichman (1986) who takes a discourse analysis approach to multimodal dialogues. The second component of a step describes the possible and actual ways to implement the corresponding dialogue step. It may be implemented by a single dialogue act. In this case the variety is provided by the different forms the utterance may have. For instance, the presentation of a certain set of data may take the form of a list of the data records, a table, or a graphical presentation. However, the step may also be a certain sequence of dialogue acts which then build a sub-dialogue that may – in accordance with the COR model – replace the single act. Thus, we have a means to prescribe a certain act as appropriate in the given situation, which allows the user to perform it in a way she prefers, e.g. by requesting context or help information.

An approach to problem solving based on past experiences is pursued in the area of case-based reasoning (CBR). In our experimental work, we adapted the ideas of CBR to the requirements of a user-guidance component (for details cf. Tißen 1991, 1993, Stein, Thiel & Tißen 1992) which was developed as part of a prototypical information system.

An Application – the MERIT System

MERIT (Multimedia Extensions of Retrieval Interaction Tools) is a prototypical knowledge-based user interface to a relational database on European research projects and their funding programs (cf. Stein, Thiel & Tißen 1992). The database contains textual and factual information (a subset of the CORDIS databases which are offered online by the ECHO host). These data were extended by interactive maps and scanned-in documents and pictures the user may request as additional context information in certain situations (cf. for example fig. 5 below). The system features form-based query formulation (various form sheets for different 'perspectives' on the data), and the visualization of retrieval results in different situation-dependent graphical presentation forms. One major system component is a case-based dialogue manager (CADI, cf. Tißen 1991) which controls the retrieval dialogue and provides a library of "cases" stored after previous sessions with MERIT. The cases are used to guide the user through the current session, proposing thematic progression basically by suggesting a specific order of query and result steps focusing on a specific perspective in each step.

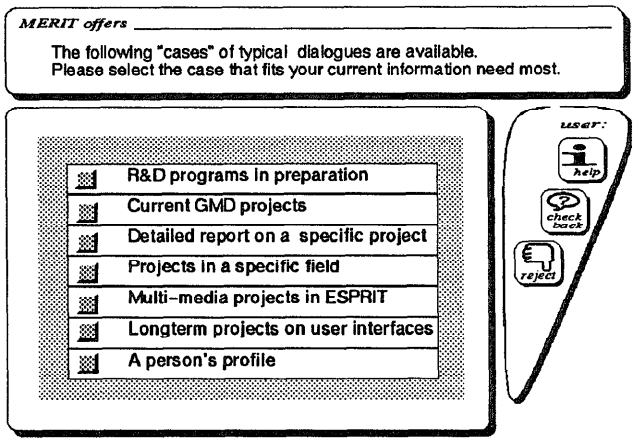

Figure 4: A system's 'offer'

The example in figure 4 shows an offer made at the beginning of a dialogue session. Here the user is asked to choose the case that suits her information need best. The graphical presentation of complex dialogue contributions – like offers, requests, inform acts – is composed of several distinct elements: the label in the upper left corner identifies the contributor and dialogue act type (e.g. "MERIT offers", "user requests", "MERIT presents"). A short text, mostly elliptic, summarizes the content of the dialogue act or gives some meta-information ("Please select ..."). Further, the concrete proposition is displayed below (here in the form of several alternatives among which the user may choose). The right bar is reserved for icons representing possible (local) user actions that refer to the current system's contribution. The user may, for instance, *reject* the offer, *request help*, or pose clarifying questions (*checkback-icon*).

A Coherent Interaction Model of MERIT

Like most advanced graphical interfaces, MERIT offers a wide variety of dialogue control options to the user (cf. icons in fig. 4 and fig. 5). In a given situation, the user may proceed in the current case, start sub-dialogues within the current step, switch to another case, etc. Usually, in conventional interfaces dialogue options are presented to the user as additional components of the interface. However, such additions are disturbing in a direct manipulation interface, since they require context-dependent method evaluation.

In the following, we outline how our conversational model allows us to integrate even complex meta-dialogic options into a coherent interpretation of the multimodal interaction.

Local, i.e. case- or situation-dependent, options are:

help, check back: From the conversational perspective the user engages in a sub-dialogue related to the system's current dialogue contribution (soliciting context information). Thus, the parameter setting of the meta-function is determined in a natural way.

reject offer, withdraw ..., continue: The meaning of these icons is intuitively grasped. They comply with transitions in the basic COR schema (cf. fig. 1). For instance, 'continue' would lead to state <1>, whereupon the user either formu-

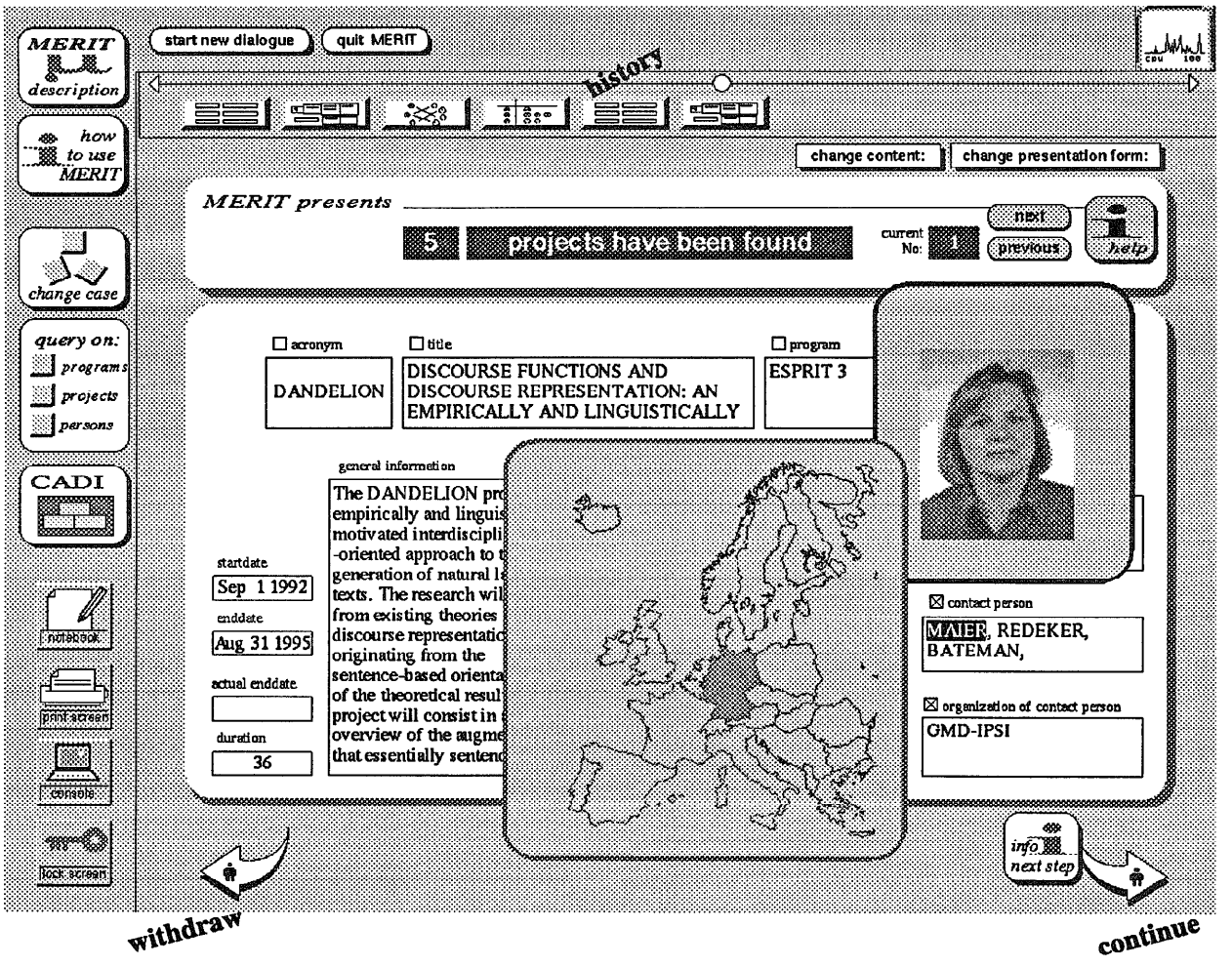

Figure 5: Example screen of MERIT with dialogue control objects

lates a new request (a query), or the system comes up with a new offer and/or information (presentation of data).

change content/ presentation: Here the user can enter a sub-dialogue related to the system's inform act and request a paraphrase and/or solicit context information (cf. fig. 2). The available options in a given dialogue state are, for instance, to ask for more detailed information about the currently presented items, to restrict the presentation to a subset of objects and attributes, or simply to replace a given presentation form by another one (e.g. a table by a graph).

The following actions are interpreted as user requests that initiate inserted meta-dialogues:

info on next step: Before the user decides whether to continue she may click on this icon and start a meta-dialogue about the system's strategy. MERIT generates situation-dependent information (a text), describing the next and subsequent steps proposed in this situation by the current case.

history: By clicking one of the history icons the user starts a short meta-dialogue referring to inform or query states of previous dialogue cycles. The respective information (her query or the retrieved data) will be displayed. The user can then compare it to the current state and decide whether to return to the current state or to go back in the history.

query on ...: At any time the user has the opportunity to insert a short retrieval dialogue. She can pose a query and inspect the retrieved data, then return to the current path/ case.

change case: The user finishes her current path, returns to the top-level dialogue (state <1>) and initiates a new dialogue cycle to choose a new case.

Conclusions

The outlined comprehensive model of multimodal interaction is based on an extended notion of *dialogue acts* which can be realized by linguistic or non-linguistic means. The structure of multimodal interaction is considered under local as well as global aspects. A comprehensive set of constraints in terms of a recursive transition network describes all (local) patterns of exchange which can occur during interaction. The (global) topical structure of the dialogue is defined according

to a selected information seeking strategy. The model was applied to design and implement the MERIT system, and led to a reduction of the complexity of the user interface while preserving most of the useful, but sometimes confusing, dialogue options of advanced direct manipulation interfaces. The conversational approach permits dialogue features to be handled in an *integrative manner*, but not as separated extensions such as undo, history, and help functions.

References

Arens, Y. & Hovy, E. 1990. How to Describe What? Towards a Theory of Modality Utilization. In: *Proc. of the 12th Annual Conference of the Cognitive Science Society*, 487-494. Hillsdale, NJ: Erlbaum.

Belkin, N.J., Cool, C., Stein, A. & Thiel, U. 1993. Scripts for Information Seeking Strategies. Paper presented at: *AAAI Spring Symposium '93 on Case-Based Reasoning and Information Retrieval, Stanford University, CA, March 23-25.*

Belkin, N.J., Marchetti, P.G. & Cool, C. 1993. BRAQUE: Design of an Interface to Support User Interaction in Information Retrieval. *Information Processing & Management. Special Issue on Hypertext* 29(4) (in press).

Fawcett, R.P., van der Mije, A. & van Wissen, C. 1988. Towards a Systemic Flowchart Model for Discourse. In: Fawcett, R.P. & Young, D. (eds.): *New Developments in Systemic Linguistics. Vol. 2*, 116-143. London: Pinter.

Fawcett, R.P. & Davies, B. 1992. Monologue as Turn in Interactive Discourse: Towards an Integration of Exchange Structure and Rhetorical Structure Theory. In: *Proc. of the 6th International Workshop on Natural Language Generation, Trento, Italy*, 151-166. Berlin: Springer.

Feiner, S.K. & McKeown, K.R. 1990. Coordinating Text and Graphics in Explanation Generation. In: *Proc. of the 8th National Conference on Artificial Intelligence, Vol. I*, 442-449. Menlo Park: AAAI Press / MIT Press.

Halliday, M.A.K. 1984. Language as Code and Language as Behaviour: A Systemic-Functional Interpretation of the Nature and Ontogenesis of Dialogue. In: Fawcett, R.P. et al. (eds.): *The Semiotic of Culture and Language. Vol. 1*, 3-35. London: Pinter.

Hutchins, E.L., Hollan, J.D. & Norman, D. 1986. Direct Manipulation Interfaces. In: Norman, D.A. & Draper, S.A. (eds.): *User Centered System Design: New Perspectives on Human-Computer Interaction*, 87-124. Hillsdale, NJ: Erlbaum.

Maier, E. & Hovy, E. 1993. Organising Discourse Structure Relations Using Metafunctions. In: Horacek, H. & Zock, M. (eds.): *New Concepts in Natural Language Processing*, 69-86. London: Pinter.

Maier, E. & Sitter, S. 1992. An Extension of Rhetorical Structure Theory for the Treatment of Retrieval Dialogues. In: *Proc. of the 14th Annual Conference of the Cognitive Science Society, Bloomington, Indiana*, 968-973. Hillsdale, NJ: Erlbaum.

Mann, W.C. & Thompson, S.A. 1987. Rhetorical Structure Theory: A Theory of Text Organization. In: Polanyi, L. (ed.): *Discourse Structure*. Norwood, NJ: Ablex.

Maybury, M. 1991. Planning Multimedia Explanations Using Communicative Acts. In: *Proc. of the 9th National Conference on Artificial Intelligence, Anaheim, CA.*

McCoy, K.F. 1986. The ROMPER System: Responding to Object-Related Misconceptions Using Perspective. In: *Proc. of the 24th Annual Meeting of the Association for Computational Linguistics, New York.*

Oei, S., Smit, R., Schreinemakers, J., Marinos, L. & Sirks, J. 1992. The Presentation Manager, A Method for Task-Driven Concept Presentation. In: Neumann, B. (ed.): *Proc. of the European Conference on Artificial Intelligence*, 774-775. Chichester: John Wiley.

Reichman, R. 1986. Communication Paradigms for a Window System. In: Norman, D.A. & Draper, S.A. (eds.): *User Centered System Design: New Perspectives on Human-Computer Interaction*, 285-313. Hillsdale, NJ: Erlbaum.

Reichman, R. 1989. Integrated Interfaces Based on a Theory of Context and Goal Tracking. In: Taylor, M.M., Neel, F. & Bouwhuis, D.G. (eds.): *The Structure of Multimodal Dialogue*, 209-228. Amsterdam: North-Holland.

Rist, T. & André, E. 1992. From Presentation Tasks to Pictures: Towards a Computational Approach to Graphics Design. In: Neumann, B. (ed.): *Proc. of the European Conference on Artificial Intelligence*, 765-768. Chichester: John Wiley.

Schank, R. & Abelson, R. 1977. *Scripts, Plans, Goals and Understanding*. Hillsdale, NJ: Erlbaum.

Searle, J.R. 1975. Indirect Speech Acts. In: Davidson, D. & Harman, G. (eds.): *The Logic of Grammar*, 59-82. Encino, CA: Dickinson Publishing Co.

Searle, J.R. & Vandervcken, D. 1985. *Foundations of Illocutionary Logic*. Cambridge, GB: Cambridge University Press.

Sitter, S. & Stein, A. 1992. Modeling the Illocutionary Aspects of Information-Seeking Dialogues. *Information Processing & Management* 28(2):165-180.

Stein, A., Thiel, U. & Tißen, A. 1992. Knowledge-Based Control of Visual Dialogues in Information Systems. In: Catarci, T., Costabile, M.F. & Levialdi, S. (eds.): *Proc. of the 1st International Workshop on Advanced Visual Interfaces, Rome, Italy*, 138-155. Singapore: World Scientific Press.

Tißen, A. 1991. A Case-Based Architecture for a Dialogue Manager for Information-Seeking Processes. In: A. Bookstein et al. (eds.): *Proc. of the 14th Annual International Conference on Research and Development in Information Retrieval, Chicago*, 152-161. New York: ACM Press.

Tißen, A. 1993. Knowledge Bases for User Guidance in Information Seeking Dialogues. In: Wayne, D.G. et al. (eds.): *Proc. of the 1993 International Workshop on Intelligent User Interfaces, Orlando, FL*, 149-156. New York: ACM Press.

Winograd, T. & Flores, F. 1986. *Understanding Computers and Cognition*. Norwood, NJ: Ablex.