Making Systems Sensitive to the User's Changing Resource Limitations

Anthony Jameson,* Ralph Schäfer, Thomas Weis, André Berthold, Thomas Weyrath
Department of Computer Science, University of Saarbrücken
P.O. Box 15 11 50, 66041 Saarbrücken, Germany

Abstract:

Situationally determined limitations in users' "resources" (e.g., time and working memory) constitute an increasingly important challenge to adaptive interfaces—one which does not yield easily to straightforward solutions. This article gives an overview of a research program that emphasizes empirically based understanding of this problem and the use of an explicit model of the relevant causal relationships. After introducing the challenge and comparing several possible approaches to it, we summarize related work on adaptive systems and the empirical research that forms the basis of the READY prototype. The structure and workings of this prototype are discussed and illustrated with examples. We conclude by summarizing how the results obtained so far can form the basis for practically applied systems.

Keywords: Adaptive systems, User modeling, Bayesian networks, Time pressure, Working memory, Speech, Dialog

1 Introduction

1.1 *Situational Factors as a Challenge for Intelligent User Interface Design*

The guest editors of the special issue of this journal devoted to IUI98 called attention to the increasing importance of the *context* of computing as an issue in human-computer interaction: "As users carry computers, in various forms, around with them the context of their computing is varied and varying" ([17], p. 1).

Their comments referred to research ([5]) that concerned the user's software context—i.e., the other programs that the user is working with at the same time as a given target system. An even more salient aspect of context is what's going on in the world around the user. Figure 1 illustrates the type of interaction that is becoming more and more typical of the way in which users deal with computing devices. For some time now, computer users have been able gather the information they need to plan and prepare trips in the relatively tranquil setting of their home or office. Nowadays they can increasingly do so during the trip itself. The two users in Figure 1 have just flown into an airport and are using their PDAs to look for the next train connection that will get them from the airport train station to their next destination. The

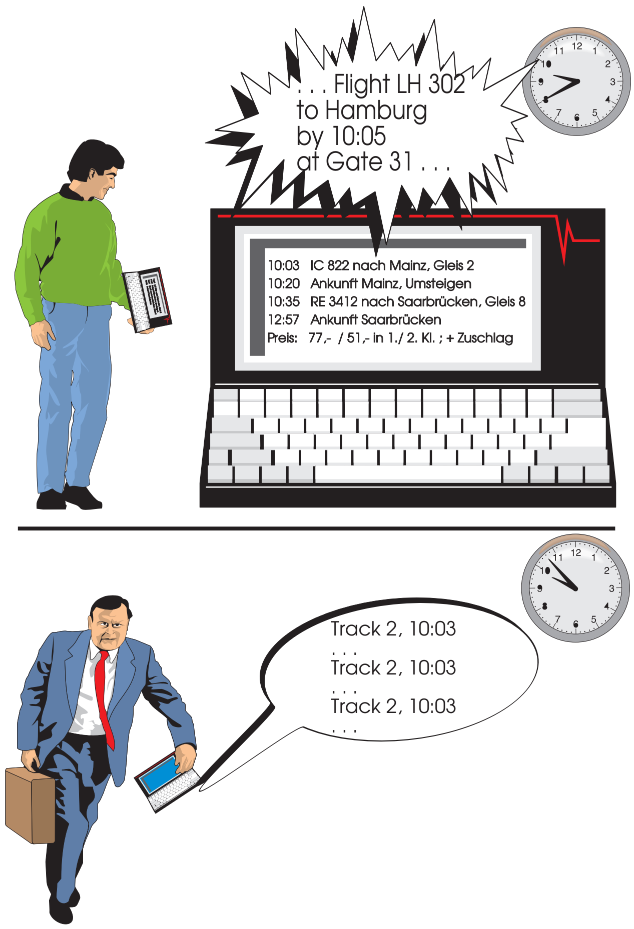

Fig. 1. Examples of interaction in which differing resource situations call for different system behavior.

effects of context on the second user are especially clear: Since the next train will be leaving in a few minutes, he has minimal time to study the information presented by the system. Moreover, he will not be able to concentrate fully on this information while navigating within the airport to the station.

The user depicted at the top of the figure has more time until his train leaves, so he is not subject to much time pressure or distraction from concurrent activity. But even he has to deal with a situational distraction—the loudspeaker announcements—which can interfere with his processing of

*Corresponding author. Tel.: +49 681 302-2474; fax: +49 681 302-4136; e-mail: jameson@cs.uni-sb.de

the information presented by the system.

It is sometimes useful to view situational factors as creating temporary limitations on the *resources* that the user can make use of while interacting with a system. In our research, we focus on the resources of (a) time and (b) working memory—i.e., the temporary storage of factual and control information that is required for the performance of one's current task(s).

1.2 How Complex Does a Solution Have to Be?

Even if we grant that situationally determined resource limitations need to be taken into account in the design of interactive systems, it is not obvious that intelligent, adaptive user interfaces are needed for this purpose. And even within the community of researchers on intelligent user interfaces, there is a growing consensus ([29]) that it's best to try to get by with as little intelligence as possible: Start with the simplest solution possible and see how far it gets you; improve it incrementally by dealing with the problems that it fails to solve. Accordingly, we should start by considering several general approaches to the problem of situational resource limitations, starting with the simplest one.

1. When designing, just assume minimal user resources

Designers often know in advance that the application (or device) that they are designing will typically be used in situations where the user's resources will be limited. Accordingly, many fixed aspects of the design—such its degree of complexity—take these limitations into account. But even for a particular application on a particular device, there can be a good deal of variability in the users' resources, as Figure 1 illustrates. System behavior that is optimal for one case may be inappropriate for others. For example, the first user in Figure 1 would probably not be satisfied with the minimal information given to the second user; on the other hand, he would have more difficulty than the second user in dealing with speech output.

2. Allow users to specify appropriate system behavior

A somewhat more ambitious approach would involve *adaptability*: Given that the users know what situation they are in, why not let them specify the type of system behavior that's most appropriate, as is done in Figure 2? Although this approach has been applied successfully in many contexts, it suffers from limitations that are especially serious in the present context:

1. The user has know that such options are available and remember to specify them at the appropriate time.
   Achieving this goal may be fairly easy if the user is familiar with the system. For novice or infrequent users, the system would have to advertise the options more or less obtrusively, using up some of the resources it was trying to save.
2. Making the appropriate choices places demands on user resources.
   Aside from the time physically required to make the necessary selections, the user may need some thought to figure out what type of system behavior is best for his or her current situation. Moreover, as we will see below, the implications of situational resource limitations for appropriate system behavior are often quite subtle. So no matter how many resources might be devoted to the task, the user may not make appropriate choices.

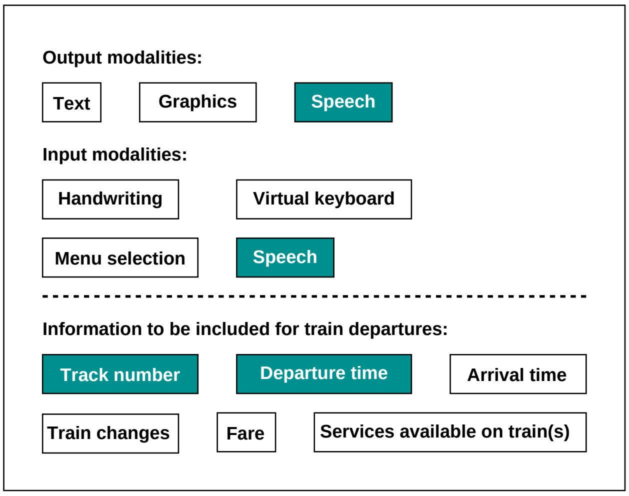

Fig. 2. Dialog box through which users could specify system behavior appropriate to their current resource limitations.

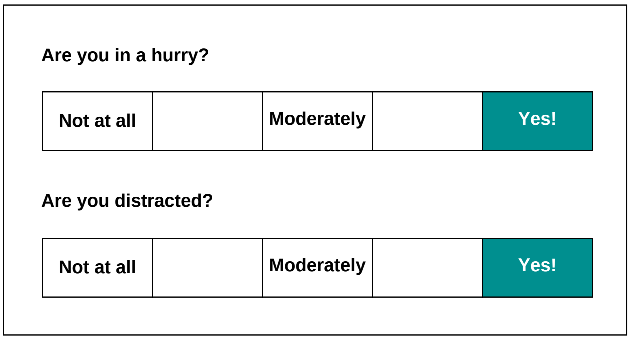

Fig. 3. Dialog box through which users could characterize their own resource limitations.

3. Have users characterize their own resource limitations

Maybe users can at least give some indication of their momentary resource limitations, leaving it to the system to decide how to adapt to them, as in Figure 3. This approach shares the first limitation listed for the previous approach. Moreover, even if successful, this approach handles only half of the problem, that of *recognizing* resource limitations; the problem of *adapting* to them still remains.

4. Have the system adapt to aspects of the user's behavior according to straightforward principles

It is actually very easy to design a system that adapts noticeably to a user's resource limitations. Consider, for example, the following two rules, which refer to a user $\mathcal{U}$ and a system $\mathcal{S}$:

1. "If $\mathcal{U}$ talks fast
   then $\mathcal{S}$ should synthesize fast speech"

2. "If $\mathcal{U}$ asks for clarification of $\mathcal{S}$'s output
   then $\mathcal{S}$ should simplify subsequent outputs"

Rules like this have a certain plausibility, and their use would make it unnecessary for the system to reason about unobservable variables such as the user's subjective time pressure or mental load. But on closer inspection it becomes questionable whether they would lead to the desired results. For example, Rule 1 presupposes (a) that fast talking by $\mathcal{U}$ indicates that $\mathcal{U}$ is in a hurry and (b) if $\mathcal{U}$ is in a hurry $\mathcal{U}$ will be helped if $\mathcal{S}$'s speech output is relatively fast. In reality, of course, there are various reasons why someone may speak unusually fast; and fast speech output may end up wasting $\mathcal{U}$'s time rather than saving it (for example, if $\mathcal{U}$ is unable to process the rapid speech correctly). In short, there is a complex web of causal relationships that determine $\mathcal{U}$'s observable behavior and the consequences of $\mathcal{S}$'s behavior for $\mathcal{U}$. So even if some simple strategies like these could be useful, it is not easy to determine which ones these are.

1.3 *The Approach Taken Here*

This article summarizes an approach to this problem that emphasizes empirically based understanding and explicit modeling.

In the next section, we summarize research on adaptive systems that has some bearing on this problem. We then summarize the four types of empirical research that we have been conducting. We then motivate and present a decision-theoretic framework for representing and reasoning about these relationships. The prototype dialog system READY shows concretely how this framework can be realized and the type of adaptation that can be achieved.

2 Relevant Previous Work on Adaptive Systems

2.1 *Time Limitations*

An especially explicit type of adaptation to situational time constraints is exhibited by the text generation system PAULINE (see, e.g., [10]), which also takes into account a number of other rhetorical goals. The approach exemplified by READY builds on PAULINE's approach in two ways: (a) Instead of simplifying its own processing when time limitations are perceived, READY focuses on predicting and controlling the processing time of the user—which seems likely to become a more important factor in the long term. (b) Instead of employing heuristic rules to adapt to time constraints, it uses explicit causal models of the relationships among the relevant variables.

A method for explicitly anticipating the user's processing of a presentation in a time-critical situation was presented in [8] (see also [7]). This decision-theoretic approach takes into account, for example, the facts that (a) presenting additional information to an equipment operator in an emergency may lead to a better decision by the operator, but (b) the utility of that decision may be lower because of the additional time required for the system to transmit and the operator to process the additional information. This work illustrates how a decision-theoretic framework supports the formalization of the tradeoffs that have to be dealt with when scarce resources have to be allocated.

The problem of how a system can *recognize* the time constraints of a user in the first place has not to our knowledge been addressed yet.

2.2 *Working Memory Limitations*

Several adaptive systems have modeled aspects of the user's cognition that are closely related to working memory (WM). For example, the tutoring system of [19] includes a method for computing the *cognitive load* imposed on the student by a specific type of task, as well as procedures for imposing the optimal level of load in each individual case. One of the user models employed in the LUMIÈRE prototype, (see [9]) included an unobservable hypothesis USER DISTRACTED as well as assumptions about its relationship to the difficulty of $\mathcal{U}$'s current task and to $\mathcal{U}$'s observable behavior.

These systems deal with links between WM-related variables and specific other variables. The examples suggest that it would be worthwhile to develop a more general conceptualization that can be applied in a wider variety of situations.

3 Empirical Foundations

When thinking about empirical research related to intelligent user interfaces, people tend to think first of *evaluations* of a system's effectiveness. And indeed, the ultimate goal is to produce demonstrably effective systems. But as was argued in Section 1.2, we need a good deal of new insight in order to be able in the present context to create a prototype that is worth evaluating in the first place. The empirical evaluation of a system that was not already based on empirical research would be a time-consuming exercise that would most likely reveal that the system was not particularly effective, without helping us to understand why not.

Although global system evaluation is planned for a later stage of research on READY, up to now we have been building up an empirical basis in four other ways:

Caller:	<groan>
Fireman:	Fire Department.
Caller:	Yes, this is Frau Schmidt. Schopenhauer Street 10. My water heater is on fire. Just this morning this repairman was here <loud breathing>. Now the thing's on fire. <loud breathing>.
Fireman:	Schopenhauer Street 10?
Caller:	Schopenhauer Street 10.
Fireman:	In what part of town?
Caller:	Pardon?
Fireman:	In what part of town?
Caller:	Well, that's here in Saarbrücken <pause>, in Sankt Johann, near the LVA.
Fireman:	What floor?
Caller:	<incomprehensible>
Fireman:	So what floor is that?
Caller:	Uh, the first. <groan>
Fireman:	Close, close, ... <pause>
Caller:	<softly:> Oh hurry! <loud breathing>
Fireman:	Yeah, close the door, we're coming right away, OK?
Caller:	Yes, that's OK.
Fireman:	Yes, bye.

Fig. 4. Translated transcript of an emergency call to a fire department.

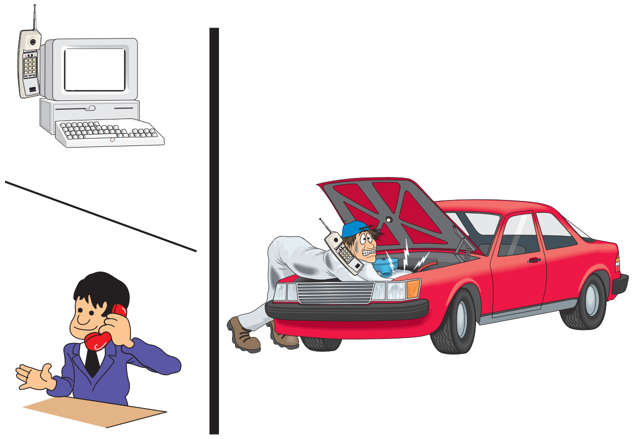

Fig. 5. Illustration of READY's initial example scenario, showing the similarity of READY's role to that of an auto repairman who offers advice by phone.

3.1 *Knowledge Elicitation from Persons with Relevant Experience*

In connection with spoken dialog as a communication medium, there exist people who have some expertise in recognizing and adapting to other people's time and WM constraints: those who regularly handle emergency telephone calls. The insights provided by such persons can usefully complement those obtained by other means, especially when they concern variables and relationships that are hard to capture objectively.

For example, ten firemen from the Saarbrücken Fire Department served as subjects in a study ([30]) that made use of the method of *retrospective thinking aloud* (see, e.g., [6]). Each subject listened to three previously recorded phone calls from persons reporting fires. The tape was stopped at predetermined points. At each such point, the subject was asked to answer spontaneously a question about the caller that presumably corresponded to a type of assessment that firemen make while handling such calls (e.g., "How quickly will she be able to provide the information that she was just asked for?"). The subject was then asked to verbalize the thoughts that occurred to him while answering the question.

An analysis of the answers yielded a picture of the (largely shared) causal relationships that the subjects implicitly perceived. Some of these relationships were consistent with general results from previous experimental research (cf. Section 3.3 below), while others were fairly domain-specific (e.g., the link between audible breathing and agitation, which in turn is seen as a cause of "lack of concentration").

In addition to specific qualitative causal relationships, the analysis yielded the following general conclusions:

1. Firemen handling emergency calls perceive varying degrees of "concentration" in the callers, and they believe that these variations should be adapted to. For example, several subjects noted that it would have been pointless to give the caller in Figure 4 further instructions after dispatching the fire engine, since the agitated caller would probably have had difficulty remembering and executing them anyway.

2. Most of the subjects' formulations included expressions of uncertainty and vagueness (e.g., "…seems to be pretty distracted …"). The firemen seem to recognize implicitly that the relevant causal relationships are probabilistic in nature and do not permit reliable, precise assessments or predictions.

3. An important source of error is a failure to assess accurately the *demands* that a specific dialog contribution will place on the listener. For example, most of the subjects were initially surprised at the difficulty that the caller in Figure 4 had in stating "what part of town" her street was in (though they were able to offer post hoc explanations of the difficulty of this question).

This last result illustrates that, even where there is converging evidence that yields a fairly clear assessment of $\mathcal{U}$'s resource limitations, this assessment is only one step toward successful adaptation. Methods for accurately predicting the resource demands of particular system behaviors are equally important.

3.2 *Qualitative and Quantitative Analysis of Data from Real Interactions*

It is often useful to have detailed, objective—and where possible quantitative—data on what goes on in the sort of interaction that one is interested in. For this purpose, knowledge elicitation can be usefully supplemented by the analysis of detailed records of real interactions.

We conducted a study of this sort in the context of the first example scenario of the READY prototype, which is illustrated in Figure 5: Users are drivers whose cars need minor repairs; they request assistance from the system in natural language via mobile phone.

In our field study, this scenario was realistically realized with the help of a professional auto repairman and 9 naive subjects, who dealt with intentionally created auto repair problems. The transliterated dialogs confirmed that in this scenario drivers sometimes show signs of WM overload (for example when trying to perform an unfamiliar task with the car while talking to the repairman on the mobile phone). In this study, time pressure was manipulated with differing deadlines and monetary rewards.

The speech of the subjects and the mechanic was analyzed in detail (see [3], [4]). The analysis showed for example, which features of speech occurred frequently enough—and with enough variability—to be potentially useful as symptoms of resource limitations.

On a more qualitative level, the input that we give to READY when testing the system—and the dialog contributions that it is designed to be able to produce—are modeled closely on the transliterations of these dialogs. In this way we hope to ensure that READY provides realistic solutions to realistic problems, in spite of its status as a research prototype.

3.3 Synthesis of Results from Previous Laboratory Research

Even the most detailed observational data usually tell us little about *causal* relationships—especially when these involve variables, such as time pressure and cognitive load, that are difficult or impossible to observe. Hence the importance of experiments in which variables are systematically manipulated. A lot of relevant results can be found in the published experimental literature from various fields, even though the experiments were in almost all cases conducted for different purposes.

Consider, for example, Oviatt's research on the prospects and problems of spoken language as a medium for human-computer communication (see, e.g., [22]). From our perspective, it yields quantitative data about some *causes* of high WM load in users (i.e., the need to supply lengthy and/or unstructured speech input) as well as some observable *consequences* of high WM load (i.e., filled pauses, self-corrections).

Other results relevant to speech—which so far has been the interaction medium of the READY prototype—come largely from psycholinguistic experiments. For example, Roßnagel ([24]) created high WM load in one group of speakers by forcing them to retrieve a large amount of information from long-term memory while speaking; observable consequences included a large proportion of pauses and a diminished quality of the content of the utterances produced.

Experimental results concerning the consequences of time limitations for speech production are less numerous and less conclusive. For example, some studies have shown that speakers under time pressure sometimes speak faster and make more self-corrections ([20]) and produce less task-unrelated information ([25]); but these results applied only under certain conditions.

Our distillation and (partly quantitative) integration of relevant results of previous empirical research ([3], [4], [12]) has proven useful as a foundation for the development of the READY prototype, even though a good deal of extrapolation is in general required to bridge the gap between the original experimental situations and the system's actual application scenarios. In particular, [4] illustrates how analyses of this type can be combined with analyses of observational data for the derivation of qualitative and quantitative constraints on implemented models of users' resource limitations.

3.4 New Laboratory Research

The utility of previous laboratory research for the design of a particular type of intelligent system is typically limited in two ways:

1. Some important causal relationships have received little or no attention, since they were not of interest to previous researchers.
2. The quantitative results are typically not reported with the degree of detail that one would like to have when implementing a model that is capable of making inferences about specific cases.

New experimental studies can be designed to fill at least some of these gaps. For example, in a recent experiment ([31]) we compared the effectiveness of two ways of providing a sequence of verbal instructions ("stepwise" vs. "bundled") as a function of (a) the length of the sequence and (b) whether or not the user had to perform a secondary task while following the instructions.[1] To supplement the traditional statistical analyses of the data of this experiment, we are currently applying techniques for the learning of Bayesian networks to the data, so that the probabilities in the networks will reflect the regularities uncovered (see also [13]).

In sum, each of these types of empirical research contributes in a different way to our understanding of the causal relationships. The next section will consider how the results of such research can be represented and exploited in an adaptive system.

4 Choice of a Suitable Modeling Framework

How can we design an intelligent interactive system so that it can recognize and adapt to the user's changing resource limitations? Research on user and student modeling has yielded a variety of techniques for assessing and adapting to properties of users, including logic- and stereotype-

[1] The experiment was designed and conducted in collaboration with Leonie March and Ralf Rummer of the Department of Psychology, University of Saarbrücken.

based techniques, machine learning methods, and a host of qualitative and quantitative application-specific procedures.[2]

For the problem at hand here, what seems most suitable is a decision-theoretic framework that includes *dynamic Bayesian networks* and closely related *influence diagrams* for modeling the user's resource limitations and making decisions about the system's behavior.[3]

For the *recognition problem*, these methods are especially well suited for

1. integration of unreliable evidence from a diverse set of observations (in particular, concerning causes and symptoms of resource limitations);
2. incremental use of sparse evidence (as opposed, e.g., to the processing of large amounts of evidence with machine learning methods); and
3. explicit reasoning about the ways in which the user's resource limitations change during the interaction.

For the *adaptation problem*, these methods allow

1. exploitation of most parts of the same probabilistic model that is used for the recognition task;
2. comparative evaluation of possible system behaviors using multiple evaluation criteria whose weights depend on $\mathcal{U}$'s resource limitations; and
3. consideration in the evaluation process of predicted user responses to system behaviors.

Before demonstrating these properties of the framework with reference to the prototype system READY, we will introduce the context and overall architecture of this system.

5 Overview of the READY Prototype

5.1 Example Dialog

The example dialog in Figure 6 was conducted with the prototype with a view to illustrating the main features of the way in which the system works. For simplicity, in each dialog turn both $\mathcal{S}$ and $\mathcal{U}$ are forced to choose one of just two possible utterances: a relatively verbose one (shown on the left) and a more compact one.

The overall course of the dialog can be summarized as follows: Initially $\mathcal{S}$ expects a low level of time pressure and a fairly high available WM capacity. $\mathcal{U}$'s first utterance (2) provides no reason to change these expectations, but his problem description (4) is unusually terse and to the point. $\mathcal{S}$ therefore assesses him as probably being in a hurry and not suffering from much cognitive load. $\mathcal{S}$'s question (5) and instruction (7) accordingly emphasize conciseness at the expense of additional information that would make it easier for $\mathcal{U}$ to understand and respond appropriately.

Once $\mathcal{U}$ starts looking under the hood of the car (8), he begins producing utterances that suggest an increase in cognitive load: The repetition of part of the instruction ("Cooling water filter ...") is typical of someone who is having a hard time maintaining all of the necessary information in working memory (and untypical of someone who wants to get the dialog over with as quickly as possible). $\mathcal{U}$'s subsequent request (9) for $\mathcal{S}$ to repeat the instruction reinforces this impression. Note that these changes in $\mathcal{S}$'s assessments don't imply that $\mathcal{S}$'s earlier assessments were inaccurate: Perhaps $\mathcal{U}$ only began experiencing high cognitive load once he looked under the hood (e.g., because of lack of familiarity with the things he saw there). This example illustrates how a user's resource limitations can change within the course of an interaction, forcing the system to track a moving target.

$\mathcal{S}$'s subsequent utterances are accordingly more explicit than $\mathcal{S}$'s earlier ones, requiring less thought and search on $\mathcal{U}$'s part.

5.2 Form of Interaction With the Prototype

READY's user interface is designed so as to make it unnecessary for us to deal with the challenging problems of speech processing raised by the scenario, allowing us to concentrate on the causes and consequences of various features of the behavior of the user and the system. Input is done via a natural language menu interface with which the "user" can compose utterances and specify a number of aspects of their form, such as the position and length of pauses (see Figure 8). As Figure 7 shows, the system's output is likewise presented as text.

5.3 Architecture

Figure 9 shows the current system architecture. What is sent to the DIALOG MANAGEMENT component is a representation of $\mathcal{U}$'s utterance that contains the information that READY needs in order to update its user model and determine an appropriate response, i.e.: the meaning of $\mathcal{U}$'s utterance in the current dialog context plus any evidence that can be extracted from the input (including noises from the environment) that bears on $\mathcal{U}$'s current resource limitations.

The DIALOG MANAGEMENT component uses its knowledge about the current dialog state and about possible diagnosis and repair actions to determine a set of possible dialog contributions (e.g., instructions) that might make sense in the current situation—not yet taking into account $\mathcal{U}$'s resource limitations.[4] These contributions may differ in their basic content (e.g., prescribing a simple or a complex action) and/or in their form (e.g., using simple, redundant formulations or concise, technical ones). The DIALOG MANAGEMENT component then sends the possible contributions to the USER

[2] The volume [14] includes a representative sample of these techniques, and its Reader's Guide provides a classification.

[3] The classic reference for this family of techniques is [23]. A survey of their use for user and student modeling is given in [11], while a briefer more recent discussion is offered in [9].

[4] As with the processing of input utterances, $\mathcal{S}$'s knowledge about dialog structures and auto repair problems has been kept as simple as possible, so that we can focus on the key problem of adapting to $\mathcal{U}$'s resource limitations.

1.	*S*	Good morning to you. You are connected to the automatic repair advice service READY of the SFB378.	**Good morning, automatic repair advice service READY of the SFB378.**
2.	*U*	Good day to you, too. This is Joseph Smith from Eckerhausen.	**Joseph Smith here.**
3.	*S*	Mr. Smith, how can I help you? Are you having problems with your engine?	**What kind of problem do you have?**
4.	*U*	<Um> The ... lamp under <eh> next to the meter for the cooling water is lit up red.	**Cooling water warning lamp is on.**
5.	*S*	Have you noticed anything else unusual about the cooling system, maybe steam around the grill at the front of the hood?	**Anything else unusual about the cooling system?**
6.	*U*	At the front of the car−<uh> engine, there's, there's something steaming there ... and the fan is off.	**The cooler is ... steaming and the fan isn't working.**
7.	*S*	Under the hood, on the right, in the front, near the yellow oil measuring stick, there's a cable. Please check its plug.	**Check the plug on the cooling water filter.**
8.	*U*	**Cooling water filter ...**	OK.
9.	*U*	**What was that again?**	Got it.
10.	*S*	**Under the hood, on the right, in the front, near the yellow oil measuring stick, there's a cable. Please check its plug.**	Check the plug on the cooling water filter.
11.	*U*	Plug next to the oil stick.	**OK.**
12.	*U*	What was that again?	**Got it.**
13.	*S*	**From the plug, there's a cable that goes to the fan. Is the cable damaged somehow−−a bit loose or bent?**	Is the connecting cable damaged?

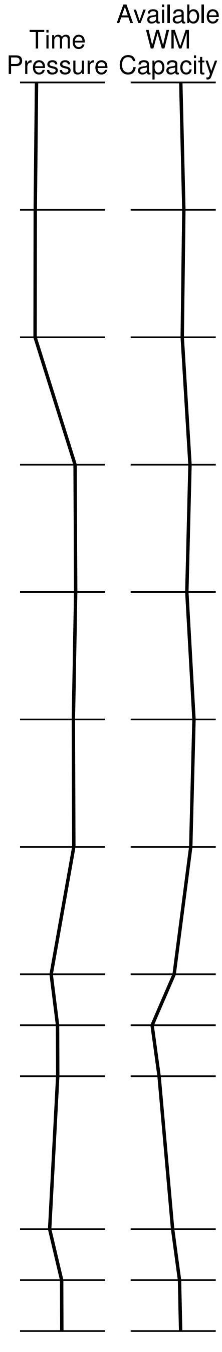

Fig. 6. Part of an example dialog with the READY prototype, translated from the original German.

(Of the the two possible utterances for each dialog turn, the one actually chosen in this dialog is shown in bold type. The graphs on the right show the changes in the expected values of $\mathcal{S}$'s beliefs concerning the two key unobservable variables..)

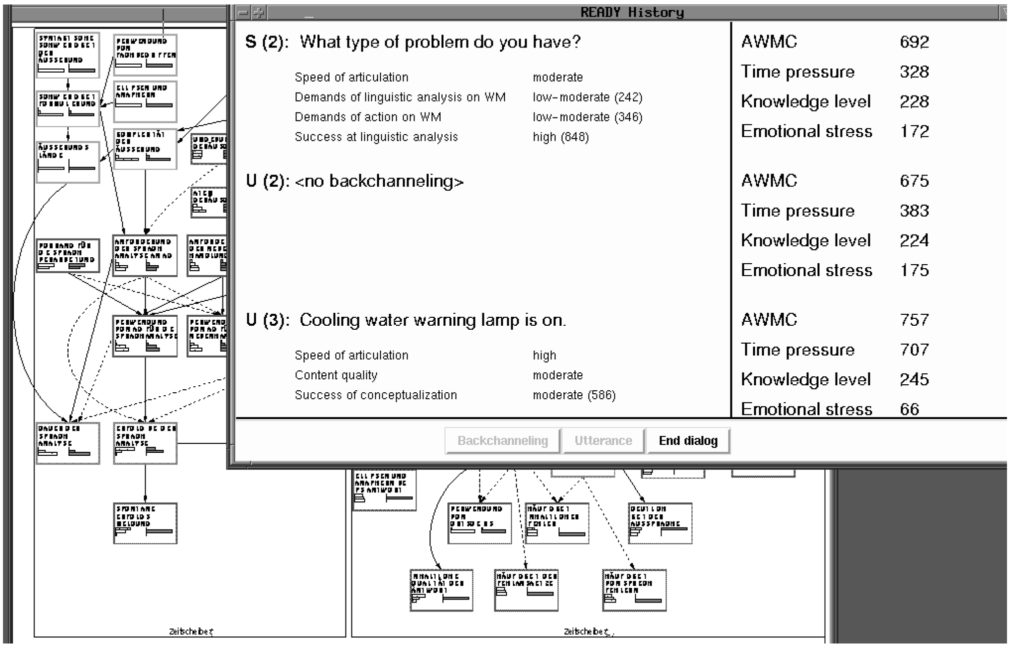

Fig. 7. Translated version of READY's main interaction screen.

(The annotations below each utterance of $\mathcal{S}$ summarize its relevant features that are not reflected in the text and $\mathcal{S}$'s assessments of various properties of the utterances that are relevant for $\mathcal{S}$'s inferences. The numbers on the right are the expected values of the system's current assessments of key variables, including available WM capacity ("AWMC")—cf. the two variables represented graphically in Figure 6. The reasons for changes in these assessments can be examined at any time in the Bayesian network for the current pair of time slices, shown here in the background—cf. Section 6.)

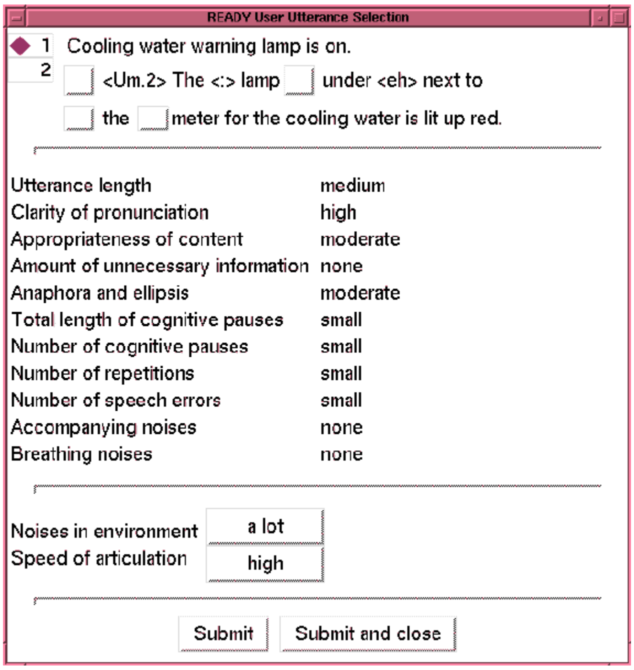

Fig. 8. Translated version of READY's menu interface for simulating speech input.

(Either of the two utterances shown at the top can be selected by the user for input; for the second, longer one, a number of variants can be specified via the pull-down menus attached to parts of the utterance. The table in the middle shows the characteristics of the currently selected utterance—here, the first one—that the system will treat as significant.)

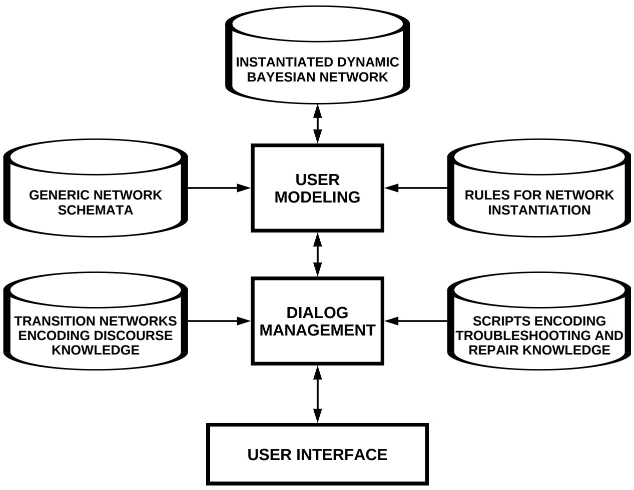

Fig. 9. Architecture of the Ready prototype.

(Boxes denote processing components, cylinders denote knowledge bases, and arrows show the flow of information.)

MODELING component, which decides which one seems best in the light of $\mathcal{U}$'s current resource limitations.

6 Modeling of Resource Limitations

6.1 *Example of a Network Schema*

All of the USER MODELING component's assumptions about relevant causal relationships are represented in *network*

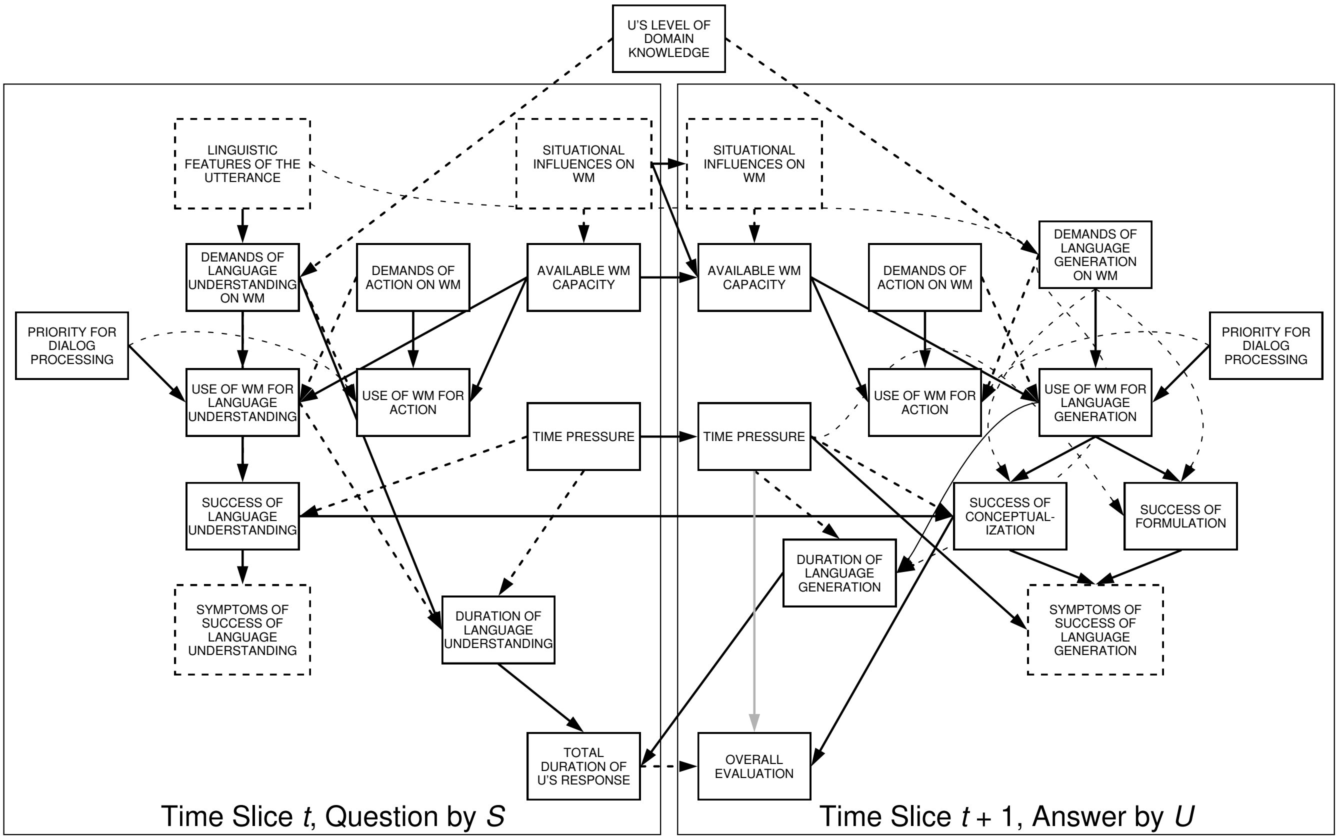

Fig. 10. A schema for the construction of two time slices of a dynamic Bayesian network.

(Solid and dashed arrows denote positive and negative causal influences, respectively. Each box with a solid border represents a node that corresponds to a single variable. Each box with a dashed border denotes a group of variables that play a similar role in the network. Solid and dashed arrows denote positive and negative causal influences, respectively..)

schemata (cf. the knowledge source in the upper left of Figure 9). Each time the USER MODELING component is asked by the DIALOG MANAGEMENT component to evaluate a possible utterance (e.g., a question), it uses the appropriate schema in order to extend its Bayesian network model of $\mathcal{U}$. Each such schema corresponds to a typical sequence of two dialog moves. For example, Figure 10 shows, in simplified form, the schema for the sequence "Question (by system) $\rightarrow$ Answer (by user)". This schema corresponds to a pair of time slices of a dynamic Bayesian network.

The variables in the left-hand and right-hand time slices are related to $\mathcal{U}$'s success in understanding and answering a question, respectively. In each time slice, there is a group of observable variables (SYMPTOMS OF SUCCESS OF LANGUAGE UNDERSTANDING and SYMPTOMS OF SUCCESS OF LANGUAGE GENERATION, respectively, at the bottom in the figure) that serve, among other things, as symptoms of $\mathcal{U}$'s resource limitations. The only other variables whose values are ever known to $\mathcal{S}$ with certainty are the ones in the group LINGUISTIC FEATURES OF THE UTTERANCE in the upper left; these represent properties of a possible utterance by $\mathcal{S}$ that will influence the demands that the utterance will place on $\mathcal{U}$'s time and WM.

6.2 *Conceptualization of Resource Limitations*

The variables in the middle level of Figure 10, such as TIME PRESSURE and AVAILABLE WM CAPACITY, correspond to theoretical constructs. Since their definitions are not tightly constrained by empirical evidence, they require some theoretical assumptions to be made; note that alternative assumptions could be introduced without changing the basic approach.

Time Pressure

The variable TIME PRESSURE is viewed as the subjective cost of time to $\mathcal{U}$. It is assumed that an increase in time pressure increases $\mathcal{U}$'s desire to reduce the duration of an action, even when such a reduction exacts a price in terms of other criteria (such as the likelihood of success of the action). For example, in the first time slice TIME PRESSURE is assumed to have a negative influence on both the time that $\mathcal{U}$ spends trying to understand $\mathcal{S}$'s utterance and the likelihood that $\mathcal{U}$ will succeed. In the second time slice, similar influences are shown for the tasks of conceptualizing and formulating an utterance.

Working Memory

In research in cognitive psychology, some highly differentiated models of working memory and its use have been developed which include several subsystems such as a *central executive*, a *phonological loop*, a *visuo-spatial scratchpad*, and sometimes further subsystems (see, e.g., [2]). For an initial effort at on-line modeling of a user's interaction with a interactive system, these models are too fine-grained. In the present version of READY's model, therefore, WM is treated as if it were a homogeneous store with a particular *capacity*. (A comparable simplification is employed, e.g., in [18].) In the near future we will systematically investigate more differentiated modeling schemas for WM to see if they contribute added value, at least for particular types of tasks ([13]).

The simplified conception is based on the following assumptions: At each point in the interaction, $\mathcal{U}$ has some *available WM capacity* that can be used to handle his or her tasks. This available capacity may be less than $\mathcal{U}$'s total WM capacity, for example if $\mathcal{U}$ is agitated or distracted by events in the environment (cf. the group of variables SITUATIONAL INFLUENCES ON WM near the top of Figure 10).

The WM capacity that $\mathcal{U}$ can devote to interaction with $\mathcal{S}$ may be reduced further if $\mathcal{U}$ tries to perform another task simultaneously. For example, in the first time slice in Figure 10, $\mathcal{U}$ might be performing an action while listening to a question asked by $\mathcal{S}$. Each task that $\mathcal{U}$ performs is assumed to create a particular *demand* on $\mathcal{U}$'s WM. When two tasks have to be performed simultaneously, the available WM capacity may be too low to permit both to be performed without problems. In this case, the way $\mathcal{U}$ uses his or her WM for the two tasks is assumed to depend on the relative priority of the system-related task for $\mathcal{U}$. This relative priority is represented in each half of Figure 10 by a variable PRIORITY FOR DIALOG PROCESSING.

6.3 Temporal Relationships Among Variables

As was mentioned earlier, one of the challenges involved in the modeling of time and WM constraints is the need to deal with different types of change in the variables being modeled.

Most of the variables in Figure 10 are *temporary*: Because they refer to a brief event or state, they are defined only within a single time slice. Other variables—for example, U'S LEVEL OF DOMAIN KNOWLEDGE, are *static*: They are defined in all time slices, and any changes in their value during the course of an interaction are assumed to be negligible. But the critical variables TIME PRESSURE and AVAILABLE WM CAPACITY are *dynamic*: They are defined in all time slices, but their values can change significantly during an interaction, for example because of situational influences. The schema in Figure 10 illustrates a frequently applied method for handling a dynamic variable in a Bayesian network: A separate node for the variable is included in each time slice and is linked to the corresponding node in the previous time slice. A detailed discussion and justification of the way this method is applied in READY is given in [27] and [26], and a different application of this basic method to a user modeling problem is presented in [1].[5]

6.4 Selection of the System's Dialog Contributions

When the USER MODELING component is asked by the DIALOG MANAGEMENT component to evaluate a possible utterance of $\mathcal{S}$, it proceeds as follows: A network schema is chosen that will allow $\mathcal{S}$ to anticipate and evaluate the consequences of the utterance (e.g., if the utterance is a question, the schema in Figure 10). The Bayesian network that has been constructed so far is extended with two new time slices. The variables in the group LINGUISTIC FEATURES OF THE UTTERANCE are instantiated as dictated by the properties of the candidate utterance. The other root nodes are instantiated on the basis of information that $\mathcal{S}$ has about the current situation—for example, concerning the likelihood that $\mathcal{U}$ is (still) occupied with some physical action. The extended network is then evaluated, giving rise to new probabilistic expectations about the variables in both time slices. That is, $\mathcal{S}$ anticipates both the immediate processing of its utterance by $\mathcal{U}$ and $\mathcal{U}$'s subsequent responses to the utterance (e.g., $\mathcal{U}$'s answering of the question or $\mathcal{U}$'s carrying out of an instruction).

The purpose of this anticipation is to generate an OVERALL EVALUATION of the candidate utterance (cf. the bottom nodes in Figure 10) using evaluation criteria that are presumably in line with $\mathcal{U}$'s own interests: the total time required for $\mathcal{U}$ to respond (weighted negatively) and the success of $\mathcal{U}$'s response. The relative weight that $\mathcal{U}$ would attach to these two criteria is assumed to depend on $\mathcal{U}$'s time pressure, as is indicated in the figure.

The overall pattern is that $\mathcal{S}$'s perception of $\mathcal{U}$'s resource limitations can affect $\mathcal{S}$'s choice of utterances in two ways:

- Most directly: High time pressure biases the overall evaluation in favor of utterances that $\mathcal{U}$ can deal with quickly.
- Less directly: $\mathcal{U}$'s time pressure and WM availability can determine the speed and success of $\mathcal{U}$'s response in more or less complex ways, thus indirectly influencing the overall evaluation.

Use of Influence Diagrams to Narrow the Search

This procedure of systematically evaluating each possible system utterance can be quite time-consuming if there are a lot of possible utterances. We have therefore tested an alternative approach, which will be integrated into the prototype: Instead of creating a separate possible extension of

[5] In [9], an alternative approach is discussed that eliminates the need to create multiple time slices, while leaving implicit some of the relationships that are expressed explicitly with time slices. In a dialog system, it is useful to introduce at least one time slice for each dialog contribution, because of the qualitatively different events that are involved in each contribution.

its basic network for each possible utterance, $\mathcal{S}$ creates a single *influence diagram* (see, e.g., [28], [23]), in which several aspects of an utterance (e.g., syntactic complexity and use of technical terms) are represented by *decision nodes* and the OVERALL EVALUATION is treated as a *value node*. The procedure used for processing influence diagrams (see [16]) gives values for the variables in the decision nodes that would lead to the best possible overall evaluation—i.e., desirable properties of $\mathcal{S}$'s next utterance. It is relatively easy to find the candidate utterances that come closest to having these properties.

6.5 Updating the User Model

The utility of the procedures described in the previous subsection of course depends largely on the accuracy of the user model that has been built up in the course of the dialog. This model is updated on the basis of information about $\mathcal{U}$'s behavior that is sent to the USER MODELING component by the DIALOG MANAGEMENT component. This information may concern either $\mathcal{U}$'s immediate feedback to $\mathcal{S}$'s utterance (e.g., one of the possible SYMPTOMS OF SUCCESS OF LANGUAGE UNDERSTANDING in the first time slice) or $\mathcal{U}$'s response to the utterance (e.g., one or more SYMPTOMS OF SUCCESS OF LANGUAGE GENERATION in the second time slice).[6] The network is reevaluated, leading to new assessments of the variables in both time slices. Where static or dynamic (as opposed to temporary) variables are affected, these reassessments will influence $\mathcal{S}$'s choice of subsequent utterances.

7 Technical Feasibility

The explicit representation of many types of causal relationships makes READY's adaptation relatively easy to understand, criticize, and improve. But it does lead to a fairly high degree of complexity. For example, the two time slices of a network constructed on the basis of the schema in Figure 10 currently comprise about 60 nodes; and the complete network that is built up during the dialog includes one time slice for each dialog contribution.

The main implementation language for READY is Lucid COMMON LISP, but the Bayesian networks and influence diagrams are processed with the tool NETICA ([21]). On a SUN Ultra 1 with 147 MHz and 256 MByte of RAM, the time required for interpretation of the evidence in a single user utterance varies from about .5 sec to about 3 sec, depending on how many time slices have been added to the network; a maximum of 12 time slices are retained, any earlier ones being eliminated through a *rolling-up* procedure. Evaluation of a single possible system utterance is slower, taking up to 12 sec under the same conditions. Fortunately, the processing of a related influence diagram, which (as mentioned above) can replace the evaluation of a number of similar utterances, takes roughly the same amount of time as the evaluation of a single utterance.

[6] The utterances 8 and 9 in the example dialog in Figure 6 illustrate how both types of feedback influence the system's subsequent behavior.

In sum, the techniques as currently implemented are about two orders of magnitude too slow for practical application.

8 Conclusions and Future Developments

8.1 Conclusions

Although the research discussed here is being continued, some general conclusions can be derived from the arguments and the system discussion presented above:

1. In some cases, it is worthwhile to design systems so that they can recognize and/or adapt to users' changing resource limitations.
2. There are reasons to base the design on an explicit model of the causes and consequences of such resource limitations.
3. An empirical basis for such a causal model can be derived from various complementary types of empirical studies.
4. Dynamic Bayesian networks are well suited for the representation and use of such a causal model.
5. An implemented prototype based on these principles exhibits the the various types of reasoning and decision making that one desires in such a system.

In addition many specific results concerning particular variables and implementation techniques were derived in specific studies. Only a few examples were mentioned in this overview article; more are given in the papers cited.

8.2 Current Research

Our research is currently being expanded along several dimensions (see [13] for an overview).

One current research focus is on the improvement of the empirical basis of READY's causal model: As was sketched in Section 3.4, we are conducting experiments whose data will not only clarify important causal relationships but also serve as input to Bayesian networks learning techniques, so that the necessary conditional probabilities can be derived directly from the data.

An adaptation of the prototype to the scenario illustrated in Figure 1 and to different interaction techniques (e.g., gestural input and graphical output) will provide evidence concerning the generalizability of the approach, as well as yielding specific results that are relevant to interaction media other than speech.

8.3 Simplification for Real Use

Ready is intended as a prototype that can serve as a starting point for the development of deployable systems. Here is a list of possible simplifications, starting with the least drastic ones:

1. Simplification of the network structure at points where it proves to be unnecessarily differentiated in practice.
   For example, a variable for which the belief shows neg-

ligible changes in practice can be eliminated; or a group of variables (such as the ones represented by the dashed boxes in Figure 10) can be replaced by a single coarse-grained variable.

2. Use of self-assessments as sketched in Figure 3, if users can reasonably be expected to be able to provide them.
The user's self-assessments could be used to instantiate observable variables in the network, either supplementing or replacing the observable variables that refer to naturally occurring aspects of the user's behavior.
Even more simply, these self-assessments could be taken at face value and used by $\mathcal{S}$ as estimates of $\mathcal{U}$'s resource limitations.
3. Use of heuristic rules as sketched in Section 1.2
We noted above that such rules are of questionable value as a starting point, because of the difficulty of formulating rules that have a reasonable chance of working effectively. But after experience has been gained with a more articulated, explicit representation such as READY's, it may be possible to identify some recurrent patterns of successful adaptation and to formulate these as rules.
4. Doing without adaptation altogether
With some types of system or domain, it may turn out that even an elaborate adaptive system ends up showing roughly the same behavior regardless of the user's momentary resource limitations. This result could come about for various reasons: There might seldom be enough evidence available during system use to discriminate different levels of resource limitation; there might exist particular system behaviors which are better than the alternatives no matter what the user's resource limitations are.
Note that even in a case like this, the design of a system would have benefited from the analysis of the way the system should behave given a variety of possible resource limitations.

9 Acknowledgements

This research is being supported by the German Science Foundation (DFG) in its Collaborative Research Center on Resource-Adaptive Cognitive Processes, SFB 378, Project B2, READY. We thank the Saarbrücken Fire Department for its cooperation and Wolfgang Wahlster for his continuing support.

This article is a revised and expanded version of [15]. The revision benefited from suggestions made by the four IUI 99 reviewers and from comments of the conference participants.

References

[1] D. W. Albrecht, I. Zukerman, and A. E. Nicholson. Bayesian models for keyhole plan recognition in an adventure game. *User Modeling and User-Adapted Interaction*, 8:5–47, 1998.

[2] A. D. Baddeley. *Working Memory*. Oxford University Press, Oxford, 1986.

[3] A. Berthold. Repräsentation und Verarbeitung sprachlicher Indikatoren für kognitive Ressourcenbeschränkungen [Representation and processing of linguistic indicators of cognitive resource limitations]. Master's thesis, Department of Computer Science, University of Saarbrücken, Germany, 1998.

[4] A. Berthold and A. Jameson. Interpreting symptoms of cognitive load in speech input. In J. Kay, editor, *User Modeling: Proceedings of the Seventh International Conference, UM99*, Vienna, New York, 1999. Springer Wien New York.

[5] A. K. Dey, G. D. Abowd, and A. Wood. Cyberdesk: A framework for providing self-integrating context-aware services. *Knowledge-Based Systems*, 11:3–13, 1998.

[6] K. A. Ericsson and H. A. Simon. *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA, Rev. edition, 1993.

[7] E. Horvitz. Transmission and display of information for time-critical decisions. Technical Report MSR-TR-95-13, Microsoft Research, February 1995.

[8] E. Horvitz and M. Barry. Display of information for time-critical decision making. In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 296–305. Morgan Kaufmann, San Francisco, 1995.

[9] E. Horvitz, J. Breese, D. Heckerman, D. Hovel, and K. Rommelse. The Lumière project: Bayesian user modeling for inferring the goals and needs of software users. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 256–265, Madison, WI, 1998.

[10] E. H. Hovy. *Generating Natural Language Under Pragmatic Constraints*. Erlbaum, Hillsdale, NJ, 1988.

[11] A. Jameson. Numerical uncertainty management in user and student modeling: An overview of systems and issues. *User Modeling and User-Adapted Interaction*, 5:193–251, 1996.

[12] A. Jameson. Wie gehen wir mit dem Arbeitsgedächtnis unserer Dialogpartner um? Eine Integration von Ergebnissen aus vier Forschungsrichtungen [How do we deal with the working memory of our dialog partners? An integration of results from four areas of research]. In H. Mandl, editor, *Bericht über den 40. Kongreß der Deutschen Gesellschaft für Psychologie in München 1996*, pages 258–263. Hogrefe, Göttingen, Germany, 1997.

[13] A. Jameson. Adapting to the user's time and working memory limitations: New directions of research. In U. J. Timm and M. Rössel, editors, *ABIS-98, Adaptivität und Benutzermodellierung in interaktiven Softwaresystemen [ABIS-98, Adaptivity and User Modeling in Interactive Software Systems]*. FORWISS, Erlangen, Germany, 1998.

[14] A. Jameson, C. Paris, and C. Tasso, editors. *User Modeling: Proceedings of the Sixth International Conference, UM97*. Springer Wien New York, Vienna, New York, 1997.

[15] A. Jameson, R. Schäfer, T. Weis, A. Berthold, and T. Weyrath. Making systems sensitive to the user's time and working memory constraints. In M. T. Maybury, editor, *IUI99: International Conference on Intelligent User Interfaces*, pages 79–86. ACM, New York, 1999.

[16] F. Jensen, F. V. Jensen, and S. L. Dittmer. From influence diagrams to junction trees. In R. Lopez de Mantaras and D. Poole, editors, *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 367–373. Morgan Kaufmann, San Francisco, 1994.

[17] P. Johnson, L. Terveen, and J. Marks. Guest editorial. *Knowledge-Based Systems*, 11:1–2, 1998.

[18] M. A. Just and P. A. Carpenter. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99:122–149, 1992.

[19] A. Kashihara, T. Hirashima, and J. Toyoda. A cognitive load application in tutoring. *User Modeling and User-Adapted Interaction*, 4:279–303, 1995.

[20] E. Marx. *Über die Wirkung von Zeitdruck auf Sprachproduktionsprozesse [The Effect of Time Pressure on Speech Production Processes]*. PhD thesis, University of Münster, Germany, 1984.

[21] Norsys Software Corp., Vancouver, BC, Canada. *Netica API Programmer's Library Reference Manual*, 1997.

[22] S. Oviatt. Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction*, 12:93–129, 1997.

[23] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[24] C. Roßnagel. Kognitive Belastung und Hörerorientierung beim monologischen Instruieren [Cognitive load and listener orientation in instruction monologues]. *Zeitschrift für Experimentelle Psychologie*, 42:94–110, 1995.

[25] R. Rummer. Verarbeitungsbeschränkungen bei der Sprachproduktion: Ein experimenteller Ansatz zur Erforschung sprachlicher Makroplanungsprozesse [Processing constraints in speech generation: An experimental approach to research on macroplanning processes]. In R. H. Kluwe, editor, *Strukturen und Prozesse intelligenter Systeme [Structures and Processes of Intelligent Systems]*, pages 41–63. Deutscher Universitäts-Verlag, Wiesbaden, 1997.

[26] R. Schäfer. *Benutzermodellierung mit dynamischen Bayes'schen Netzen als Grundlage adaptiver Dialogsysteme [User Modeling With Dynamic Bayesian Networks as a Foundation for Adaptive Dialog Systems]*. PhD thesis, University of Saarbrücken, Germany, 1999.

[27] R. Schäfer and T. Weyrath. Assessing temporally variable user properties with dynamic Bayesian networks. In A. Jameson, C. Paris, and C. Tasso, editors, *User Modeling: Proceedings of the Sixth International Conference, UM97*, pages 377–388. Springer Wien New York, Vienna, New York, 1997.

[28] R. D. Shachter. Evaluating influence diagrams. *Operations Research*, 34:871–882, 1986.

[29] J. Shavlik. Bridging science and applications (panel discussion). In M. T. Maybury, editor, *IUI99: International Conference on Intelligent User Interfaces*, pages 45–46. ACM, New York, 1999.

[30] T. Weyrath. Erkennung von Arbeitsgedächtnisbelastung und Zeitdruck in Dialogen – Empirie und Modellierung mit Bayesschen Netzen [Recognition of working memory load and time pressure in dialogs: Empirical studies and modeling with dynamic Bayesian networks]. Master's thesis, Department of Computer Science, University of Saarbrücken, 1998.

[31] F. Wittig, B. Großmann-Hutter, and A. Jameson. Learning Bayesian networks on the basis of data from a psychological experiment. Manuscript submitted for publication, 1999.