Robotica (1997) volume 15, pp 117–127. © 1997 Cambridge University Press

Conversational interfaces to robots

Colleen Crangle

Medical Informatics, Stanford University School of Medicine, MSOB X-215, Stanford, CA 94305-5479 (USA)

SUMMARY

There is growing interest in robots that are designed specifically to interact with people and which respond to voice commands. Very little attention has been paid, however, to the kind of verbal interaction that is possible or desirable with robots. This paper presents recent work in multimodal interfaces that addresses this question. It proposes a new form of robot-user interface, namely a *collaborative conversational interface.* This article explains what collaborative conversational interfaces are, argues for their application in robots, and presents strategies for designing good conversational interfaces. It concludes with a discussion of the particular challenges faced in designing conversational interfaces for robots.

KEYWORDS: Conversational interface; Robots; Voice commands.

1. INTRODUCTION

For a long time it was assumed that intelligent robots would be autonomous robots. Robotics research, it was thought, should aim at the design of devices that pursue their own goals without human intervention. Increasingly, however, attention is being paid to robots that are designed specifically to interact with people. These robots exhibit an intelligence that is the result of integrating human competence with machine ability. The work on supervisory robot control, for applications such as space telerobotics, exemplifies this approach.[1] However, it is in rehabilitation robotics that the most sustained effort has taken place. Robotic aids for the physically disabled capitalize on the ability of physically disabled people to specify the desired robot action and then monitor and correct the robot's performance. Users typically communicate their intention to the robot using voice commands or head or eye movements. Our own research into instructible robots has focused on ways in which any user can teach new skills to a robot using natural-language commands. Instructible robots are designed to find application as rehabilitation robots or as service robots in other industries.[2–11]

For all such robots, the design of the user interface is central. The user must be able to communicate with the robot and the robot must be able to communicate with the user if robot and user are to function effectively together. Various approaches to robot-user interface design can be found in current research.[12–19] A significant feature of this work is its embrace of multimodal interfaces, with speech input figuring prominently. Very little attention has been paid, however, to what *kind* of verbal interaction is possible or desirable with robots. Is communication in a natural-language such as English possible? Or should some subset of English be carved out as the proper language for communicating with robots? What is the proper communicative role for the robot; should it also speak? And how do user and robot manage the sustained verbal interactions required to perform interesting tasks?

In this paper, I will present recent work in multimodal interface design that addresses these questions. This work arose within medicine, but what I discuss draws on our many years of research in instructible robots. I propose a new form of robot-user interface, namely a *collaborative conversational interface.* In a collaborative conversational interface, the user and the machine engage in a conversation couched in natural spoken language. My purpose in this paper is to explain what collaborative conversational interfaces are, to argue for their use in robots, and to present strategies for designing good conversational interfaces.

I begin by reviewing in section 2 related work on robot interfaces that use speech. I also describe our earlier work on robot interfaces that led to this work. Section 3 discusses in more general terms the use of speech recognition in robot interfaces. Section 4 presents three strategies for designing collaborative conversational interfaces, using examples from our work in instructible robots. Section 5 describes the first two conversational interfaces we have implemented for applications within medicine, and discusses the current direction our work is taking. Section 6 describes specific steps that must be taken in the design of a conversational robot interface. This section concludes with a discussion of the particular challenges faced in designing conversational interfaces for robots.

2. BACKGROUND AND RELATED WORK

In considering other approaches to interactive robot control, I want to pay particular attention to the role of speech in the interfaces. First, I consider the Cambridge University Robot Language (CURL) project.[12,13] The CURL environment provides for several different kinds of interaction: direct telemanipulation of the robotic device, command-line control of the device, and the selection of predefined robot procedures. Most recently, an iconic drag-and-drop interface has been integrated with the CURL environment to provide potential users in rehabilitation robotics with an easy and intuitive way of controlling a robot. This work provides for voice control of the iconic interface, but speech interaction as such is not part of the interface design.

Another approach to giving physically disabled users

https://doi.org/10.1017/S0263574797000143 Published online by Cambridge University Press

natural control of a robot can be found in the work of Kazi et al.[14] This work (known by the acronym MUSIIC) assigns a central role to speech input in the communication. The system interprets the user's natural-language speech inputs and deictic gestures (such as pointing) using a stereo-vision system coupled with planning methods and an object-oriented knowledge base. Although sophisticated in its use of human gestures and in its integration of human competence and machine ability, this work currently deals with simple dialogue models.

The VA/Stanford rehabilitation robotics program, which has been in existence since the late 1970s, has long had a commitment to voice interfaces.[15–19] A remarkable feature of this robotics program is the extent to which prototype robots have been subjected to evaluation and clinical assessment. For example, the voice-controlled vocational robotic assistant, known as DeVAR, was placed in two settings: a classroom, for use by a programming student for four months, and an office, for use by a senior computer programmer for 18 months.

Over the years the interfaces in this program have become more task-focused than motion-focused. The most recent work within this program, for example, focused on the following high-level robot tasks:

- Getting a cup of water from a cooler and bringing it to the front of the workstation, then returning it to the cooler.
- Retrieving a mouth-stick from its storage holder and presenting it to the user.
- Operating a mechanical timer knob on a microwave.

A similar evolution has taken place in the voice interfaces being developed. In addition, researchers quite early on recognized the need for voice feedback, that is, for letting the robot speak to the user. Research was also undertaken with our own group on more complete task-oriented dialogues such as those entailed in teaching the robot a new task.[5–11] We considered the use of a rehabilitation robot in the home for activities of daily living, and in the workplace for book-handling tasks. The experimental work made use of both a fixed manipulator and a mobile robot. No sustained exploration of speech input was undertaken in these projects. All emphasis was on advanced natural-language understanding.

Conversational interfaces are a natural extension of this work. The focus now is on sustained speech interactions between user and robot, interactions that are directed towards some mutual goal. To make these sustained interactions possible, the user requires a clear and intuitive grasp of what the robot can do, and what the robot is doing at any point in the interaction. That is, some representation of the robot, its capabilities, and the task environment is needed. A conversational interface provides such representation. The representation is both verbal and visual. Spoken noun phrases refer to and semantically represent the objects in the task domain, and spoken verb phrases (and adverbial phrases) refer to and semantically represent robot actions. Furthermore, in a conversational interface, both the user and the robotic system speak. System utterances, that is, what the robot says, are crucial in structuring the task space for the user. They let the user know what the robot can do and, as the conversation progresses, what the robot is doing. They also let the user know what the robot has understood of the user's verbal command. Visual representations of the objects and actions serve to reinforce and clarify the system utterances, thereby allowing these utterances to be expressed naturally. This process is explained in some detail in section 4.

I include this discussion of representation for the following reason. A conversational interface may appear to have much in common with a command-line interface, and carry with it all the advantages and disadvantages of such an interface. There are indeed commonalities. But the representational aspect of conversational interfaces cannot be over-emphasized. It is critical in the design of a conversational interface. Unless both verbal and visual representations are an explicit part of the design, I do not believe a usable interface can be produced. The interesting point to note here is that a similar emphasis can be found in direct manipulation systems in which each significant object and robot capability is represented for the user. The work of Dallaway and colleagues makes perhaps the most direct attempt to realize the benefits of direct manipulation interfaces for robots.[12,13] The general point I want to make is that the way we currently characterize and categorize user interfaces is not adequate for the complex interactions possible between humans and robots. With speech, these interactions will grow increasingly more complex, requiring new ways of thinking about interfaces.

3. SPEECH INTERFACES: THE ROLE OF CONVERSATION

Speech-recognition technology is now mature enough that some consider it a commodity product. Off-the-shelf recognizers such as Dragon Systems' DragonDictate exist for specific tasks such as dictation in medicine or the law. And many companies provide software tools that allow you to integrate speech recognition into your application. But we have yet to see the widespread use of speech recognition in interfaces. It is now generally agreed that the greatest challenge lies in knowing how to design interfaces that exploit the available technology. I believe that if speech interfaces are to become widely accepted, a de facto interface standard—a set of conventions—will emerge, much as it did for graphical user interfaces. And I am convinced that these conventions will be couched in terms of conversation; that is, conversation will be adopted as the fundamental design unit. There is no other way. When people use language, they engage in conversation. As the Stanford psychologist Herb Clark puts it, conversation is the "fundamental site of language use."[20] A common practice in speech interface design to date has been to let users say what they see on the screen. This style of interface design is often called "voice-enabling" the graphical interface. In general, these interfaces simply do not work. They are not used. And the reason, I believe, is that conversation was never the fundamental design unit. In these interfaces, users

https://doi.org/10.1017/S0263574797000143 Published online by Cambridge University Press

are asked to utter odd bits of language, words and phrases that do not form any meaningful whole. The spoken language does not fit together in any coherent way. In addition—and though this may seem like a minor, pragmatic point, it is more important than first appears—it just does not feel right talking to something that does not talk back to you, that does not itself speak. Speech has to be both spoken and heard to work in an interface. Much of the later discussion in this paper, in section 4, will show how important bi-directional sound is to the design of a successful speech interface. In fact, once the back-and-forth so characteristic of ordinary human speech is admitted as an integral part of speech communication with robots, problems that once seemed intractable start to admit of solutions. It will develop this point in more detail in the next section.

First, let me talk more about conversations. It might seem that asking for conversations between humans and robots is like asking for a fictionalized HAL-like machine that not only recognizes what we say but understands it. On the contrary, there is every reason to believe that spoken natural language will present far fewer demands than written language when it comes to managing dialogue, that is, when it comes to coordinating an exchange of utterances directed to some mutual goal. The literature contains a wealth of information on the differences between spoken and written language. (See Wulfman et al. for an excellent discussion of what these differences mean for speech interface design.[21]) For example, in speaking as opposed to writing, people tend to use shorter words, a smaller variety of words, and less complex sentences. They also refer to themselves more and repeat themselves more. Of course, other differences between spoken and written language point to increased difficulty in correctly *recognizing* speech—spoken language has more pauses and dysfluencies, for example. But in terms of the characteristics of ordinary conversation, spoken language reveals some well-established practices that we can exploit in designing user interfaces to robots.

For example, when people give spoken instruction on how to perform a task such as assembling an object, they break the task down into smaller components.[22] They also execute strategies for making sure that the person being addressed is following the instruction. Consider this excerpt from a transcript of a conversation, for example:

S. Okay now, the small blue cap we talked about before?
J. Yeah.
S. Put that over the hole on the side of that tube
J. Yeah.
S. that is nearest to the top, or nearest to the red handle.

The speaker identifies the item that is to be manipulated, receives acknowledgement of the identification, then tells the other person what to do with the item. The speaker also modifies and clarifies the instruction after it has been given. The practice of checking mutual understanding before proceeding in a conversation has been noted by others.[21] What has also been observed is that once a name or description has been mutually acknowledged, the participants in a conversation will gradually shorten the name or description they use.

Well-established conversational practices such as these permeate our interactions with each other. There are also certain styles of conversation that develop within different cultures. The Tasmanian Peter Conrad, author of ''Where I Fell to Earth: A Life in Four Cities,'' writes that because of the vehemence with which the Portuguese speak, he had always assumed, before he learned the language, that they were constantly quarreling. The directness and forthrightness of North American speech is often startling to people from British Commonwealth countries. Furthermore, specific activities within a culture frequently give rise not only to certain styles of conversing but to exchanges that are quite severely codified. In these exchanges, few variations are tolerated. Once you ''get it,'' it's easy. But before you get it, you can be quite mystified as to what is going on. As an interesting example, consider the exchanges that take place when ordering in a fast-food restaurant. And if you want to experience for yourself the extent to which these exchanges are codified, especially in the United States, try ordering a Coca-Cola so that you get it without ice. Don't even think of asking for *A coke without ice,* or of saying *I want a coke but I don't want any ice,* or *Give me a coke but don't put any ice in it.* You will not get what you want. You have to say *Coke, no ice,* and that's it.

I am convinced that successful speech interfaces will consist of conversations that will be codified in various ways. They will not be free-ranging interchanges. They will not have in back of them extensive natural-language understanding capabilities. They will be reduced and to a certain extent they will be systematic. They will be structured by the application, that is, by the robot, its environment, and the tasks it is to perform. And they will be structured by the interface designer who will have a clear grasp of a range of ordinary conversational conventions, the kind we adhere to in our discourse with one another every day. I also believe that certain conventions will start to emerge specifically for conversing with robots, much as there are conventions now in most cultures for conversing with strangers, fast-food servers, retails clerks, bank tellers, and others.

To illustrate this last point, let me give you an example of an interesting *failure* in a speech interface. An application was designed to allow users to speak via the telephone to a computer at a major financial institution to perform various tasks such as checking account balance, moving funds from one account to another, and so on.[23] The opening line of the conversation in this interface, that is, the first thing the computer said to people who called, was *Can I help you*? And the most frequent response by puzzled callers was *Main Menu, Please.* Two things were going on here. First, the conversation opener *Can I help you*? cued users into an

https://doi.org/10.1017/S0263574797000143 Published online by Cambridge University Press

inappropriate conversational convention altogether, not the one the system was set up to handle, but one that was suggested, I should think, by the friendly bank teller of days gone by. Second, the users were themselves, in their confusion, cueing into the only computer "conversation" they knew in this context—the menu system they previously encountered when they phoned into the voice-response system and used their telephone touchpad to reply. Conversational conventions come into play in speech interfaces whether we expect them to or not.

4. COLLABORATIVE CONVERSATIONAL INTERFACES: THREE STRATEGIES

There is an important idea that has come out of studies of human language use. It is that conversation is a collaborative process.[20] The idea is that speaking and listening are parts of a collective or joint activity, much like shaking hands or paddling a two-person canoe. It takes coordination; it takes collaboration. The challenge we face in designing a speech interface is to set up and foster a collaborative conversation, that is, a conversation that seems natural enough to users that they will engage in the process collaboratively, as in a normal conversation between two people.

Before looking at specific ways to foster collaboration, it is useful to consider two pragmatic design goals that should hold for all speech interfaces to robots. First, users should know without hesitation or effort what they can say and what the robot can do at each step. It may seem obvious that users must know what they can say, but it is very hard to get this right. The more natural the communication, the more likely users will be to use natural language in all its richness, and the less likely the system will be able to recognize what the user has said. Without exception, speech recognition systems perform better when the recognition context can be limited at each stage of the conversation. There have to be restrictions on what the user can say. But if you fail to make these restrictions clear, it is like giving the users of a graphical user interface a screen the size of a football field and telling them to click somewhere in the top right-hand quadrant, and then having only a few penny-sized clickable spots in that corner that actually initiate any action. Furthermore, these restrictions have to be posed in a way that is natural to the point of transparency. It is doubtful, for instance, that users will accept (or readily understand) restrictions represented as a grammar. The restrictions may also not be so stringent that the communication departs in a significant way from natural speech. If the restrictions are too severe, what is being imposed on the user is a new semi-technical language, and one that is probably poorly defined to boot.

The second general design goal is that user utterances should be readily recognized by the system; that is, the interface should be designed to achieve maximal recognition performance. Once again, this may seem an obvious goal—who wants to talk to something that keeps on getting it wrong?—but as with the first goal, it is difficult to achieve. There have been tremendous advances in speech-recognition technology over the past decade, but it is still the case that you have to design "around the technology." Each product has its own strengths and weaknesses and a speech interface has to be designed with these in mind. The first and second goals interact in an obvious way. Well chosen restrictions on what the user can say at each point in the interaction not only guide the user, they serve to limit the recognition possibilities and so improve recognition performance.

I will show how these two pragmatic goals can be met in the design of collaborative conversations between user and robot. I start by recognizing that we have *all* of the following interface elements at our disposal. See Figure 1. In the center lies the user's inputs, namely, the utterances. Around that we have the robot's contributions to the conversation: verbal prompts (typically questions or commands), audio responses (typically statements), graphical displays, and visual responses (robot action or animation).

Each element has a purpose to serve, as shown in the right-hand column of Table I. The robot's *verbal prompts,* either questions or commands, let the user know what the system can do and what the user can say to get the robot to take action. They guide the user in her interaction with the robot. The *graphical displays* clarify and reinforce the verbal prompts. I will have much more to say about this later. The *audio responses* help carry on the robot's side of the conversation. One particular purpose they serve is to reassure the user that what she said was heard and is being acted on. *Visual responses* further show the user that what she said was heard and is being acted on. A visual response may be actual robot action, or it may take the form of animation on the screen. All these interface elements must work in concert, clarifying and reinforcing each other.

To examine this claim more closely, consider verbal prompts in the form of questions. We know from our understanding of language that questions can be categorized according to the type of answer they expect. There are three categories:

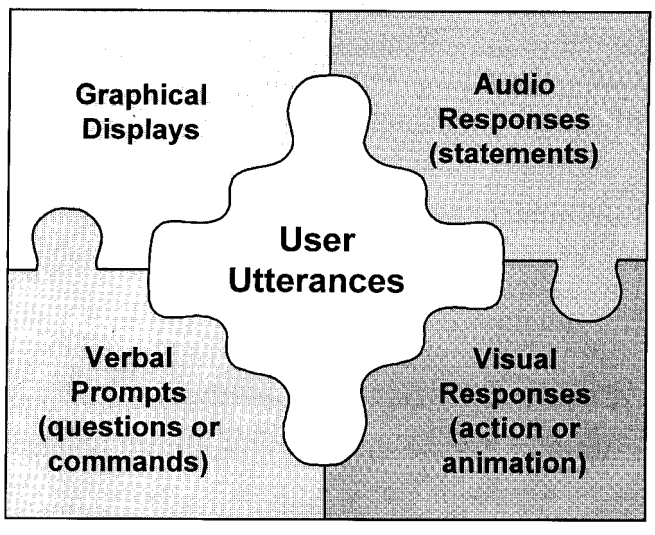


Fig. 1. Interface elements

https://doi.org/10.1017/S0263574797000143 Published online by Cambridge University Press

Table I. Interface elements and their purpose

Robot	Verbal prompts (questions or commands)	Let the user know what the system can do Let the user know what he or she can say
	Graphical displays	Reinforce the verbal prompts Clarify the verbal prompts
	Audio responses (statements)	Tell the user what the system is doing
	Visual responses (action or animation)	Show the user what the system is doing
User	Utterances	Navigate the graphical user interface, by responding to system prompts, being guided by system displays, being reassured by system responses

(i) **Questions that expect only affirmation or rejection.** These are called Yes–No questions. Examples from interaction with a robotic assistant are:

- *Is this the book you want me to get*?
- *Are you ready for me to start*?
- *Do you want a drink of water*?

Questions of this type are widely used in speech-recognition systems because Yes and No have been recognizable for some time now across populations of speakers. These questions have a completely natural role to play in conversation but are very limiting if they are the sole mode of interaction. Yes–No questions can be further categorized as to the grammatical *form* of the question, and in consequence, the expected grammatical form of the answer. Consider these three questions.

- *Do you want me to start over again*?
- *Would you like me to start over again*?
- *Should I start over again*?

A parroting effect has been noted in speech interfaces in which the user will tend to respond using the same words that occur in the prompt. So for these three questions, users will tend to respond with *Yes I do, Yes I would,* and *Yes you should,* respectively.

(ii) **Questions that expect a reply supplying an item of information.** These are called Wh-questions. Examples from interaction with a robotic assistant are:

- *Which cup do you want to use*?
- *What do you want to eat*?

Questions of type (ii) typically start with one of the interrogative words: *which, what, when, who, where, how,* and so on. The kinds of answers they expect can, in general terms, be predicted: a time (*when*), place (*where*), process (*how*), reason (*why*), object (*what, which*), person (*who*), and so on. But unless further qualifications are given in the question, a wide range of answers may be elicited. With a wide range of possible answers, it is difficult to predict what the user might say. The recognition system is then far more likely to mis-identify a word or phrase.

(iii) **Questions that expect as the reply one of the options presented in the question.** These are called Option-questions. Examples from interaction with a robotic assistant helping to prepare a meal are:

- *Do you want pasta or soup*?
- *For dessert, do you want vanilla, chocolate, strawberry or peach ice cream*?

Questions of type (iii) are in many ways optimal for speech interfaces because they present an explicit enumeration. Type-(iii) questions are an entirely natural way of making clear to the user what the expected verbal responses are and restricting the recognition context. Not only does the question make clear to the user what her options are, it does so using the specific words it hopes to get back, and syntactically encoding the form of the desired reply. However, as with Yes–No questions, excessive use of Option-questions produces a very unnatural way of communicating. Most of the time we pose open-ended questions and we rely on the context for supplying the enumeration.

This strategy is exactly the one to adopt in designing conversational interfaces. Questions of type (ii)—open-ended Wh-questions—are transformed into questions of type 3—Option-questions. This transformation is achieved by using visual displays that label the options—vanilla, chocolate, strawberry, peach, for example. The combination of the auditory prompt and the labeled display produces an enumerated list of options without sounding or feeling unnatural. In the informal usability tests we have conducted on one of the conversational interfaces described later in this paper, users have been satisfied with this form of interaction and in fact have sometimes not noticed that the information was coming from two sources—sound and text—until I pointed it out.

The first strategy for fostering collaboration in conversational interfaces, then, is just this.

Strategy 1: Use visual displays (graphics and text) to transform open-ended Wh-questions into enumerated Option-questions

The second strategy leads directly from the first. Wh-questions with labeled displays successfully substitute for enumerated Option-questions, and do so generally in a way that feels as natural as open-ended questions. However, the options are not always as neatly categorical as the ice-cream flavors of the above example, nor as easily represented in a label.

Consider the following question posed by a robotic assistant who is helping set up a workstation area: *Where do you want the multimedia speakers to go—on the shelf, the table, or next to the monitor?* The difficulty is that the concept "next to" is semantically complex. One way to display it is to depict the scene and to show one speaker on either side of the monitor, for instance. This

https://doi.org/10.1017/S0263574797000143 Published online by Cambridge University Press

Table II. Responses to *Do you want ice cream?*

	Response	Kind of Flexibility
a	Yes Yes please Yes, I do Yes, I want ice cream I do	Syntactic variation on the question
b	OK Sure I want some	Referential variety
c	Yes, I'll take vanilla	Skipping ahead
d	Yes, I like it No, ice cream is too cold Yes please	Extraneous information
e	It's my favourite food It's far too fattening	Indirect speech acts

arrangement is most likely the intended one for dual speakers and stereo sound. However, there is no one phrase that captures this configuration precisely and simply. The user might indeed respond to the display and the question *Where do you want the speakers to go?* with *Next to the monitor.* But she might just as readily say *At the monitor* or *One on each side.** The important point, however, is that there is a limit to the variations that reasonably describe the displayed option. Enumerating them is an entirely tractable problem. These variations have to be identified so that the speech-recognition system can be cued for them.

This second strategy for fostering collaboration in conversational interfaces looks at thinks from the user's point of view. Speaking on behalf of the user it says, Cut me a little slack here. Consider again the robot's asking the user what she wants for dessert, this time with the simple Yes–No question *Do you want ice cream?* And consider the five sets of responses shown in Table II.

a. The speech recognizer should be cued for straightforward *syntactic variations* of the expected reply. Many of these can be generated automatically using natural-language parsing tools.
b. Some referential variety should also be allowed, that is, different ways of referring to the same thing. *OK* and *Sure* represent different ways of saying *Yes.* The third example, *I want some,* involves anaphora, the use of a word—*some* in this case—as a grammatical substitution for a word or phrase used earlier—*ice cream.*
c. The response *Yes, I'll take vanilla* is an example of skipping ahead in the conversation, that is, anticipating the next question—*What flavor do you want*?—and answering both questions in one. Skipping ahead is a specific case of a more general conversational phenomenon that I will talk about later.
d. In ordinary conversation, responses often include extraneous information. Increasingly, commercial speech-recognition systems are providing key-word-spotting capacility in their recognizers. That is, they allow the developer to stipulate key words that the recognizer listens for, ignoring all else in the utterance. True flexibility will come when the interface developer can stipulate one or more key words to listen, plus the order of their appearance, with all else in-between being ignored. Then the *um*'s and *er*'s of ordinary speech will prove less of a problem than they do now.
e. Utterances that convey their meaning indirectly (known as indirect speech acts) represent a challenge to the design of conversational interface, and I know of no serious attempt to cue recognizers for utterances that indirectly answer question as these examples illustrate. I do not think speech communication with robots is jeopardized if such responses are not allowed. They represent a sophisticated form of interaction that is not absolutely necessary for humans and robots to carry on useful conversations.

The second strategy for fostering collaboration in conversational interfaces, then, is this.

Strategy 2: Allow reasonable flexibility in the user's responses to a prompt

Before moving onto the third and final strategy that I discuss in this paper, I want to note a point of tension between the first two strategies. The more flexibility you allow in user responses, the more you risk the user's not knowing what he or she can say. Some balance must be struck between being flexible and establishing a conventional form of interaction. As I argued earlier, human–robot conversations will become codified in certain ways, just as many of our human conversations have. We have established these conventions in human conversations because they are efficient and suit the interaction. In ordering fast food, for instance, I do not want to get to know the server or spend time on social niceties such as *How are you.* I want to get my food fast. Other interactions codify indirection, such as in *May I trouble you to move over,* or they simply codify the customary way of doing things, such as how we order items from a catalogue over the telephone. Many of our interactions with robots will be codified; it is a natural way to use language. The right kinds of balance in human–robot conversations will evolve as we gain more experience and find out how people want to talk to robots.

Let me move on to the third strategy for fostering

* There is no arguing that in fine detail these different responses mean different things. Our own work on natural-language communication with robots has explored the language of action, the interpretation of locative expressions and stop-commands, and the robot's response to natural-language corrective feedback, to name just a few complex issues.[2] But in the absence of complete natural-language understanding and robots that can reason about our intentions, communiction and action in the world of human–robot interaction will be restricted. My concern is to let those restrictions be reasonable and natural.

https://doi.org/10.1017/S0263574797000143 Published online by Cambridge University Press

collaboration in conversational interfaces. When people converse with each other, each participant makes moves to redirect the conversation. This give and take of control is a fundamental part of conversing. There are a variety of conversational gambits we employ. The two I believe we will not give up in talking with robots are these. We interrupt each other in conversation, and we change the subject.

How can we incorporate these gambits in conversational interfaces? First, consider the need to interrupt, both what is being said and what is being done. The need for interruptions is compelling. The user might change her mind as to what she wants the robot to do, or not like the way the robot is responding. Or the user may need to correct a mis-recognition; the recognizer may think she said *I want to see the graph* when what she actually said was *I want to switch off.*

It is conceptually simple to allow the user to interrupt the robot's speech. But we have faced considerable technical challenges in making that interruption work in our interfaces. In our first conversational interface, we configured our system with two sounds cards, one for user input (that is, for the exclusive use of the speech recognition system) and the other for system output (that is, for audio prompts and responses to the user). It is possible with this configuration to have the system listen at all times for the user's input and to have that input accepted, even while a prompt or message was playing. At first, our design required the user to talk over the system's audio output. In human conversations we are accustomed to having the other participant's quitting if we interrupt and persist in our interruption. When the system continued to talk, in the same tone as it had been talking before the interruption, the effect was disturbing to users. Our second approach cut off all audio output as soon as the user's utterance was recognized. Until robots function as social beings (if ever they do), it is appropriate to give the user this power to unilaterally quieten the other participant.

It goes without saying that interrupting the robot should result not only in silencing it, but also in interrupting what it was doing, if that was indeed what the interruption was about. The command *Stop pushing the chair,* for example, should not only interrupt the robot's reassuring patter that it has heard our command and is moving the chair over to the table; it should also cause the robot to stop pushing the chair. However, it is seldom a trivial matter, in any system, to allow interruptions and cancellations.[4] I mention the need to provide them, however, because once a conversational interaction is underway, it is hard to prohibit the user from saying something as natural as *Stop* or *Don't.* And with speech recognition, there is always the possibility of mis-recognition. Users must be allowed to interrupt a robot that is acting in response to a mis-recognized command.

A second essential gambit is changing the subject. This is surely the most important form of conversational redirection. Without it, the user is merely a reactive agent in a dialogue completely controlled by the robot. Earlier I referred to the user's skipping ahead in a conversation when he knows the direction it is going to take, specifically, when he thinks he knows what the next question is going to be. In interface terms, what we are talking about when we talk about skipping ahead in a conversation is *bypassing* the graphical user interface. Changing the subject is the most dramatic form of bypassing the graphical user interface. Most of what I have discussed so far concerns the way users can *navigate* the graphical user interface, responding to verbal prompts, being guided essentially by those prompts and the graphical displays that reinforce and clarify the verbal prompts. One of the chief advantages of a speech interface, however, is that it can allow the user to bypass the graphical user interface altogether. In bypassing the graphical user interface, the user takes charge and expresses his wish directly and immediately. With this form of interaction, the interface comes close to meeting the functionality of the old command-line interfaces we knew in the early days of computing and which still prevail in many robot interfaces. Command-line interfaces suffer well-known problems, but their appeal to experienced users is immense. Conversational interfaces potentially offer similar appeal.

Two challenges arise in allowing users to bypass the graphical user interface in a conversational interface. First, how do users know what to say the first few times they want to bypass the graphical user interface? And how do they remember what to say? This topic is addressed in more detail in another paper.[24] Let me summarize the main points here. The first problem—How does the user know what to say?—can be addressed by implementing a conversational convention modeled more or less on the Operator Please! or Teacher! convention, that is, a simple appeal to speak to the authority in the situation, the person who can help. On the screen, a display shows the phrase users should speak when they need help. That phrase may simply be the words *I want help.* In response, the system gives a context-sensitive verbal prompt of the following form:

> *Yes, what do you want to do? You may interrupt me at any time. Do you want to* $\langle X_1 \rangle$, $\langle X_2 \rangle \cdots$ *or* $\langle X_n \rangle$?

$\langle X_1 \rangle$, $\langle X_2 \rangle \cdots \langle X_n \rangle$ represent n of the tasks the robot is able to perform at that time, the n that the user will be told about. I have restricted n to 4 in the two conversational interfaces built to date. Empirical evidence seems to suggest that three to five options are optimal. More than that and the user does not remember what was said. It is an interesting design challenge to structure the system's knowledge of itself—what it is doing and what it can do at any time—to produce these context-sensitive prompts. Note that while the instantiation of $\langle X_1 \rangle$, $\langle X_2 \rangle \cdots \langle X_n \rangle$ changes from one occurrence of the prompt to another, the syntactic form of the question remains constant: *Do you want to . . . ?* And the words used in the instantiations of $\langle X_1 \rangle$, $\langle X_2 \rangle \cdots \langle X_n \rangle$ remain as constant as possible. For example, the prompt may be instantiated in the following ways for a robotic assistant:

https://doi.org/10.1017/S0263574797000143 Published online by Cambridge University Press

- *Do you want to eat, work at your computer, or give me some other task*?
- *Do you want to work at your computer, read a book, or give me some other task*?
- *Do you want to continue reading this page or do you want to go on to the next page*?

Using a standard form for all the context-sensitive prompts, and having them be as simple as possible, is designed to help the user learn what to say in the first place and then also to remember what to say. The intent is for the user to learn after a while, without explicit training, what the bypassing commands are. Our informal user studies so far suggest that users catch on remarkably quickly, perhaps bolstered by the fact that they can ask for a reminder at any time. These bypassing utterances are essential in a conversational interface. Without them, you do not have a speech interface; you simply have a voice-driven graphical user interface.

This, then, is the third strategy for fostering collaboration in conversation with robots.

Strategy 3: Permit users to exercise familiar conversational gambits such as interrupting and changing the subject

5. EXAMPLES OF CONVERSATIONAL INTERFACES IN MEDICINE

I now describe briefly two conversational interfaces that have been developed at Stanford University, both for applications within medicine. The first application is concerned with providing access to on-line medical information, the second with helping medical students learn anatomy. These two examples illustrate the ideas presented in the previous section.

On-line Access ot Medical Information.* There are a growing number of on-line medical information resources that contain information important for clinical practice. However, potential users face several problems when they try to use computer-based information resources to answer their questions. Each information source—whether a bibliographic citation database, a knowledge base, or an on-line medical textbook—typically has its own user interface, which the user must learn, and its own way of displaying results, which the user must interpret. A program was developed, called the Knowledge Server, that provided uniform access to several full-text on-line oncology sources. Our goal in building a conversational interface to the Knowledge Server was to help the busy clinician express his or her information need, choose an information source to look at, and then select one or more sections within a source so that the particular paragraphs of interests can be read.[25]

We made use of the speaker-adaptive, large-vocabulary speech-recognition product from Dragon Systems, called DragonDictate for Windows (DDWIN), augmented by their medical vocabulary. DDWIN requires users to pause between words when no context is given for what they are saying and their words must be identified using the 180,000-word vocabulary of the system. This mode of recognition is called ''Dictate'' mode. When the user first inputs her search concept in our interface, DDWIN must be in this mode. In addition to Dictate mode, however, DDWIN also has a ''Command'' mode in which it tracks the visible windows in the Windows application that is running concurrently with it, listening for the spoken form of captions displayed using standard Windows techniques such as option or command buttons. In this mode, the user is able to use continuous speech. In our application, programmed in Visual Basic, DDWIN tracks the captions in the displays, listening, for example, for the name of an information source or the name of a heading in one of the information sources. DDWIN also tracks captions listing the user commands for bypassing the graphical user interface; these are positioned along the lower edge of each interface display window, their content hidden to the user but ''visible'' to DDWIN's tracking mechanism. Except for the initial entry of a search concept, continuous speech is used for all user inputs in our interface. Our current implementation makes use of two sound cards: one is the IBM M-ACPA card which DragonDictate takes over exclusively for speech input; the other is the Sound Blaster 16 which the interface uses for sound output.

A typical conversation with a new user of the system runs like this. The context-sensitive system prompts that teach the user how to bypass the graphical user interface are initiated by the words *Knowledge Server, Please*.

System	*Welcome to the Oncology Knowledge Server. I will help you search several on-line information sources. Right now I'm loading the table of content for . . .*
User	*I want to start the Knowledge Server.*
System	*What search concept do you want to use? Please pause briefly between each word – like – this – and say 'Enter' when you are finished.*
User	*radical* ⟨PAUSE⟩ *mastectomy ENTER*
System	*I'm looking to see what the Knowledge Server knows about this concept.*
System	*The Knowledge Server does know this concept. Here are the frequencies of this concept's occurrence in the information sources. Blue bars show . . . Green bars show . . . Which information source do you want to look at*?
	Labeled graph is displayed.
User	*Principles and Practice*
System	*I'm finding the places where your search concept appears.*
System	*I'll show you where your search concept appears in the information sources.*

* The contributions of the following people to this application are acknowledged: Robert Carlson, Lawrence Fagan, Mark Erlbaum, Kevin Keck, Nels Olson, David Sherertz and Mark Tuttle.

https://doi.org/10.1017/S0263574797000143 Published online by Cambridge University Press

System *Here are the results. Highlighted sections show where your search concept appears. Which section do you want to look at*?

The table of contents of the textbook "Principles and Practice of Oncology" is displayed, with sections highlighted for the user.

User *Regional Nodal Involvement*

Paragraphs are displayed.

User *Knowledge Server, Please*

System *Yes, what do you want to do*? *You may interrupt me at any time. Do you want to see the table of contents, choose another information source, or enter a new search concept*?

User *I want to see the table of contents.*

The table of contents is displayed again.

User *Options in Local Treatment*

Paragraphs are displayed.

User *I want to see more.*

The user now bypasses the graphical user interface, skipping back to a previous screen and choosing another information source.

User *I want to choose another information source.*

and so on

In our most recent work on this interface, we have transformed the conversation into a three-tier structure. The first tier, suitable for the novice user, has fully informative prompts and reassuring messages at each stage of the interaction, much as shown in the above dialogue. The second tier, which comes into play immediately the user has completed her first search, reduces the length of the prompts and eliminates some of the messages. The third tier, initiated by the user's saying *I want you to be quiet* (or any reasonable variation of that request), eliminates the system prompts and messages altogether. The context-sensitive prompts initiated by *Knowledge Server, Please* always remain operative. It is this general area, namely, making the conversation adapt to the user, that is the focus of our current work in conversational interfaces.

Audible Anatomy Quiz.* Computer-aided instruction allows the teaching of anatomy to be transformed from a process of rote memorization into an interactive presentation of images and informaton accompanied by the interactive testing of the student's knowledge. We have implemented a conversational interface to an existing system, *The Anatomy Lesson,* developed at Stanford University Medical School and available commercially.[26] The conversation currently in place allows the student to test her knowledge through a quiz that solicits free-text, as opposed to multiple-choice, answers. A single, full-duplex sound card is used for both speech recognition and audio output. As with the previous interface, we have made use of DragonDictate for Windows. We are now re-doing the interface using the developer's kit from Dragon Systems. This approach will give us more ability to co-ordinate the user's and system's conversational interaction.

A typical conversation with a new user of the system runs like this, starting at the first question in the quiz. An image of a part of the body is displayed on the screen, with a flag showing what specific part of the anatomical structure is of interest. The student is typically asked to identify the structure pinpointed. Context-sensitive prompts that aim to help the user learn how to bypass the graphical user interface are initiated by the words *I want help*.

Image is displayed.

System *Please identify. Pause briefly between each word – like – this – and say 'Enter' when you are finished.*

User *I want help.*

System *Do you want a clue*?

User *Yes.*

Clue is displayed.

User *phrenic* ⟨PAUSE⟩ *nerve ENTER.*

System *That answer is correct.*

Moving on the the next question and image . . .

System *Identify. Be specific.*

User *vagus* ⟨PAUSE⟩ *artery ENTER.*

System *That answer is not correct. Do you want a clue*? *Do you want to zoom in . . .*

User *Yes.*

System *. . .on the area of interest*?

Image changes to a close up.

User *I want a clue.*

Clue is displayed.

User *I want help.*

System *Do you want the answer*?

and so on

This conversational structure has four levels, represented by the labels Talkative, Less Talkative, Quiet and Less Quiet. At the Quiet level, there is no audio output, except that initiated by the *I want help* request. At the Talkative level, the question is spoken, the evaluation (correct or incorrect) is spoken, and in response to *I want help* one or more options, depending on the context, are offered to the user. The two intermediate levels, Less Talkative and Less Quiet, adjust the amount of what is spoken by the system. Note that these levels allow the conversation to be adjusted in terms of its verbosity. There are other equally interesting kinds of adjustments that should be made in response to the user, and I am currently exploring mechanisms for implementing them and testing their effectiveness. One example is the degree of politeness exhibited by the system. Should it make requests of the user or give orders, for instance? In general, I am also considering the extent to which adjustments can and should be made automatically in response to the user's speech patterns.

* The contributions of Lin Liang and Parvati Dev to this application are acknowledged.

https://doi.org/10.1017/S0263574797000143 Published online by Cambridge University Press

6. DESIGN CHALLENGES FOR ROBOT APPLICATIONS

Conversational interfaces can be built for a wide range of applications and computer-controlled devices. I want to conclude this paper by laying out in some detail the specific steps to take in designing a conversational interface to a robot. These steps incorporate insights from the strategies described in section 4.

(i) List the tasks the robot can perform.
(ii) List the real-world objects that enter into the performance of those tasks.
(iii) Choose for each task a specific verb phrase that refers to it. This will be the standard or canonical verbal representation of that task.
(iv) Choose for each object in the robot's world a specific noun phrase that refers to it. This will be the standard or canonical verbal representation of that object.
(v) For each of the tasks identified in (i), formulate a verbal prompt of the form *Do you want to . . .* ?
(vi) For each of the tasks the user can request verbally (as defined by steps (iii) and (iv)), formulate a verbal response for the robot that will confirm what was heard and will describe what the robot is doing in response.
(vii) Organize the set of tasks identified in (i) into a hierarchical structure that represents what the robot can do at any point in its interaction with the user. For example, if one of the tasks is to prepare a meal in a microwave, there will probably be other tasks the user can refer to out of which the meal preparation task is built, such as lifting a bowl, carrying it from one point to another, and so on.
(viii) Use this hierarchy to lay out distinct recognition states for the speech recognizer.
(ix) Use the prompts in (v) to create composite prompts that fit into the hierarchy identified in (vii). These prompts should offer no more than 3–7 options at a time to the user.
(x) Devise visual displays that will allow open-ended questions to be substituted for the enumerated option prompts composed in (ix).
(xi) Identify the reasonable variations of the expected user responses to all the prompts (verbal and visual) that resulted from (x) and (xi). Use these responses to expand the recognition states identified in (viii).

These steps form the core process in the design of a conversational interface for a robot. There are many details of speech-interface design that must also be taken into account but that cannot be covered here. For example, it is important to avoid having the system listen at any one time for distinct words and phrases that sound similar enough to confuse the recognition system. Because of this constraint, it is often the case that the standard phrase for a robot action or an object cannot be the obvious choice. In general, too, longer phrases are less easily confused that one-or two-syllable words. Many of these pragmatic design decisions still hinge on the particular strengths and weaknesses of a given recognition system. Decisions have to be reached through trial and error.

I conclude this paper with some observations on two particular challenges facing the design of conversational interfaces to robots.

The first challenge arises with the application of Strategy 1: transforming open-ended Wh-question into enumerated Option-questions using graphical displays. As I remarked earlier, most of the time when we pose open-ended questions we rely on the context for supplying any implicit enumeration there might be in the question. In conversational interfaces, labeled visual displays function as the context. This works well in conversational interfaces to computers where what is on the computer screen represents the world of discourse. What is on the screen is all that there is to talk about. With robots, however, the real world intrudes. The robot and the user are physically present and that physical world is a more intrusive context than any computer display could be. Of course, the physical world is present to the robot only to the extent that it is equipped with the sensory ability to perceive it. If the robot has limited perceptual knowledge of its environment, the situation is a lot more like that with non-robotic computers. But, even so, the user will not necessarily perceive it that way. Users who are talking to a robot will be more inclined to make reference to the physical world and, however inappropriately or inadvertently, will expect the robot to know what they are talking about.

The second challenge in conversational interfaces to robots also arises from the fact that the robot and user are operating in the real world. Let me give some background to the problem. Conversational interfaces work best when in the early stages of the interaction, that is, with the novice user, the system takes the lead. This allows the system to cue the user in on what conversational conventions will come into play in the ensuing conversations. I do not yet know to what extent the leadership of the robot will work out in practice for robots designed as robotic assistants. The very concept of a robotic assistant puts the user in charge. Furthermore, and this is where the physical environment intrudes, given the richness of the environment in which many robots will be operating—the user's house or place of work, for example—the number of actions the robot can take at any time, and so the number of commands the user can give the robot at any time, will be quite large. Two potential problems exist. First, it might be difficult to restrict the recognition environment enough to achieve satisfactory recognition performance. Second, system messages to teach the user the bypassing commands will have to carefully structured so as not to overwhelm the user with options. We have already encountered this challenge in the Audible Anatomy Quiz and can see how the challenge will increase with robots.

These challenges seem well worth the effort, however.

https://doi.org/10.1017/S0263574797000143 Published online by Cambridge University Press

A robot that is designed to work in interaction with people will fully come into its own only when it participates in that most human of activities – holding conversations.

Acknowledgments

The work reported in this paper was supported in part by the following grants and resources: National Cancer Institute SBIR Contract N44-CO-33071; NIH National Action Plan on Breast Cancer 3P41 LM05305-04S2; CAMIS, the Center for Advanced Medical Informatics at Stanford; SUMMIT, the Stanford University Medical Media and Information Technologies group; ARPA-CAETI Contract N66001-95-C-8618-P00003, Lexical Technology Inc., and the Community Breast Health Project.

References

1. T.B. Sheridan, *Telerobotics, Automation, and Human Supervisory Control* (MIT Press, Cambridge, Mass., 1992).
2. C.E. Crangle and P. Suppes, *Language and Learning for Robots* (Stanford University, Stanford: CSLI Press, 1994).
3. P. Suppes and C.E. Crangle, ''Robots that learn: a test of intelligence''. *Revue Internationale de Philosophie* **44,** No. 172, 5–23 (1990). Reprinted in P. Suppes (Ed.) *Languages for Humans and Robots* (Blackwell Publishers, 1991, Cambridge, Mass.) Pp. 357–366.
4. C.E. Crangle, ''On saying 'Stop' to a robot'' *Language and Communication* **9,** No. 1, 23–33 (1989).
5. C.E. Crangle, P. Suppes and S. Michalowski, ''Types of verbal interaction with instructable robots''. In: (G. Rodriguez Ed.) *Proceedings of the Workshop on Space Telerobotics,* (JPL Publication, **II,** 87-13, 1987). (pp. 393–402 Pasadena, California: NASA Jet Propulsion Laboratory) Reprinted **In:** (P. Suppes Ed.) *Languages for Humans and Robots* (Blackwell Publishers, 1991, Cambridge, Mass.) Pp. 299–316.
6. C.E. Crangle and P. Suppes, ''Context-fixing semantics for instructable robots''. *International Journal of Man–Machine Studies* **27,** 371–400 (1987).
7. L. Liang, C.E. Crangle and L. Leifer, ''A computational model for a robotic arm instructed by natural language''. *Proceedings of the 1990 IEEE International Conference on Systems, Man, and Cybernetics.* Nov. 1990, Los Angeles, California 1990 pp. 451–456.
8. C.E. Crangle and P. Suppes, ''Instruction dialogues: teaching new skills to a robot'' In: (G. Rodriguez & H. Seraji, Eds.) *Proceedings of the NASA Conference on Space Telerobotics,* (JPL Publication 89-7, Vol. V, 1989) pp. 91–102, Pasadena, California: NASA Jet Propulsion Laboratory.
9. C.E. Crangle, L. Liang, P. Suppes and M. Barlow, ''Using English to instruct a robotic aid: an experiment in an office-like environment'' In *Proceedings of the International Conference of the Association for the Advancement of Rehabilitation Technology,* (ICAART, 1988) **Vol 3,** pp. 466–467. Washington, D.C.: RESNA—Association for the Advancement of Rehabilitation Technology.
10. S. Michalowski, C.E. Crangle and L. Liang, ''A natural-language interface to a mobile robot'' In: (G. Rodriguez, Ed.) *Proceedings of the Workshop on Space Telerobotics,* (JPL Publication 87-13, **Vol. 3,** pp. 381–392.
11. S. Michalowski, C.E. Crangle and L. Liang, ''Experimental study of a natural-language interface to an instructable robotic aid for the severely disabled'' **In:** (R.D. Steele & W. Gerrey, Eds.) *RESNA '87: Meeting the Challenge. Proceedings of the 10th Annual Conference on Rehabilitation Technology.* (Washington, D.C.: RESNA – Association for the Advancement of Rehabilitation Technology, 1987) pp. 773–775.
12. J.L. Dallaway, ''Human–computer interaction with robotic workstations'' *Proceddings of the RESNA 96 Annual Conference* (1986) pp. 339–341.
13. J.L. Dallaway, R.D. Jackson, and R.M. Mahoney, ''The user interface for interactive robotic workstations'' *Proceedings of the 1994 IEEE/RSJ International Conference on Intelligent Robots and Systems* (1994) pp. 1682–1686.
14. Z. Kazi, M. Beitler, M. Salganicoff, S. Chen, D. Chester, and R. Founds, ''Multimodally controlled intelligent assistive robot'' *Proceedings of 16th Annual RESNA Conference, Salt Lake City, Utah,* (1996) pp. 348–350.
15. L.J. Leifer, S. Michalowski, and H.F.M. Van der Loos, ''Development of an advanced robotic aid: From feasibility to utility'' *Proceedings of the Ninth Annual Conference on Rehabilitation Technology RESNA 86, Minnesota* (1986) pp. 170–172.
16. L. Leifer, ''RUI: Factoring the robot user interface''. *Proceedings of the Fifteenth Annual Conference on Rehabilitation Technology RESNA 92* (1992) pp. 580–583.
17. H.F.M. Van der Loos, ''VA/Stanford Rehabilitation Robotics R&D Program: Lessons learned in the application of robotics technology to the field of rehabilitation''. *IEEE Transactions on Rehabilitation Engineering* **3,** No. 1, 46–55 (1995).
18. H.F.M. Van der Loos, J. Hammel, ''Use of a rehabilitation robot in a classroom setting''. *Proceedings RESNA '94*, Nashville, TN (June, 1994) pp. 442–444.
19. J. Hammel, ''The role of assessment and evaluation in rehabilitation robotics R&D: Moving from concept to clinic to context''. *IEEE Transactions on Rehabilitation Engineering* **3,** No. 1, 56–61 (1994).
20. H. Clark, *Arenas of Language Use* (The University of Chicago Press, 1992).
21. C.E. Wulfman, E.A. Isaacs, B.L. Webber, and L.M. Fagan, ''Integration discontinuity: Interface users and systems''. *Proceedings of Architectures for Intelligent Interfaces: Elements and Prototypes.* Monterey, CA (March 29–April 1, 1988) pp. 57–68.
22. P. Cohen, ''The pragmatics of referring and the modality of communication'' *Computational Linguistics* **10,** 97–146 (1984).
23. S. Basson, A. Kalyanswamy, E. Man, S. Springer, and D. Yashchin, ''Establishing speech technology requirements: The Money Talks field trial'' *Proceedings of the 14th Annual International Voice Technologies Applications Conference* (*AVIOS '95*), San Jose, California (September 12–14, 1995) pp. 131–136.
24. C.E. Crangle, R.W. Carlson, L.M. Fagan, M.S. Erlbaum, D.D. Sherertz, and M.S. Tuttle, ''Conversational access to on-line medical information'' *Proceedings of the 15th International Voice Technologies Applications Conference* (*AVIOS '96*) (September 10–12, 1996, San Jose, California 1996) pp. 65–75.
25. C.E. Crangle, R.W. Carlson, L.M. Fagan, M.S. Erlbaum, N.E. Olson, D.D. Sherertz, and M.S. Tuttle, ''Meeting clinical information needs: An interface that provides uniform access across diverse information sources'' *Working Papers, AAAI-96 Spring Symposium on Artificial Intelligence in Medicine: Application of Current Technologies,* Stanford University, California (March 25–27. AAAI Press, Menlo Park, 1996).
26. P. Constantinou, S. Danne, and P. Dev, ''Transforming information for computer-aided instruction: Using a Socratic dialog method to teach gross anatomy'' *Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care* (November 1992) p. 1043.

https://doi.org/10.1017/S0263574797000143 Published online by Cambridge University Press