GENERATION OF EXPLANATIONS AND MULTI-TURN DISCOURSE STRUCTURES IN TUTORIAL DIALOGUE, BASED ON TRANSCRIPT ANALYSIS

BY

GREGORY A. SANDERS

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science
in the Graduate School of the
Illinois Institute of Technology

Approved ____________________________
Adviser

Chicago, Illinois
July, 1995

ACKNOWLEDGMENT

All theses are the labor of countless hands seen and unseen. Let us remember their toil. I acknowledge the efforts of my past teachers William Bodamer and Wesley Snyder: both were most inspiring. I am very grateful to my dear friend Lauren Campbell for keeping me going, and sane, throughout the writing of this thesis and for her love. I am grateful to my parents, Rodney and Marjory Sanders, my brothers John and Mark, and my sister Julie. No one could ask for a more wonderful family.

I have received a great deal of encouragement, advice and help from people associated with the CircSim-Tutor project. I am particularly grateful to my advisor Martha Evens for her encouragement, advice, and help with this research and for her untiring help with the administrative details of being a graduate student at IIT. I am also grateful to the other members of my committee, Peter Greene, Jim Kenevan, and Jack Snapper for their support throughout my career at IIT, and to Allen Rovick and Joel Michael of Rush Medical College for their advice and assistance with my work.

I also owe much to the encouragement and help of all my fellow students in the CircSim-Tutor project, of whom the following have each done specific work to help me with my thesis research (in alphabetical order): Stefan Brandle, Ru-Charn Chang, Reva Freedman, Michael Glass, Greg Hume, Murugan Kannan, Ramzan Ali Khuwaja, Kumar Ramachandran, Chong Woo, and Yuemei Zhang. I appreciate the time-taking assistance of those who have done classification of student initiatives using my scheme: Jafar Ali, Stefan Brandle, Lauren Campbell, Robert Reichel, and Rodney Sanders. Greg Hume was my collaborator in arriving at a consensus identification of the student initiatives in the first 28 keyboard-to-keyboard transcripts.

I appreciate the aid of Vince Carbo, Bob Ryan, and Bodhin Kjolhede during this research. They reminded me to just keep going ahead.

I am very grateful to Johanna Moore of the University of Pittsburgh and LRDC for generous advice and assistance and for allowing me to read the manuscript of her book long before it was commercially available, which enabled me to avoid trying to duplicate important work she had already done.

Finally, this work was supported by the Cognitive Science Program, Office of Naval Research under Grant No. N00014-94-1-0338, to Illinois Institute of Technology. The content does not reflect the position or policy of the government, and no official endorsement should be inferred.

G. A. S.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENT	iii
LIST OF FIGURES	vi
CHAPTER	
I. INTRODUCTION	1
Overview and Significant Problems Attacked	1
Background of Project	2
Explanation, Discourse Generation, and the Subtasks Involved	3
Discourse Generation in the Context of CircSim-Tutor	6
Contributions of this Work	8
The Text Generation vs. Cognitive Psychology Senses of the Word "Schema"	9
II. REVIEW OF THE LITERATURE	12
Text Generation as Choice or Planning	12
Text Generation from Schemas	17
Anaphoric References	27
Having a Point of View Enhances Understanding and Recall	29
Speech Acts	31
Ceteris Paribus Rules Governing Normal Conversation	31
Focus and Its Implications	35
Closely Related Work on Tutoring and Dialogue	39
Rhetorical Structure Theory (RST)	39
Alison Cawsey	41
Studies of Questions and Answers	52
Work by Johanna Moore	55
Previous Work in the CircSim-Tutor Project	57
III. STUDENT INITIATIVES	59
Introduction	59
Identification and Classification of Initiatives	63
Interesting Examples of Initiatives	68
Responding to Student Initiatives	71
Conclusion	81

CHAPTER Page

IV. ARCHITECTURAL INTEGRATION OF DISCOURSE GENERATION IN CIRCSIM-TUTOR 82

Semantic Forms 83

V. GENERATION OF EXPLANATIONS, SUMMARIES, AND MULTI-TURN STRUCTURES IN TUTORIAL DIALOGUE . 89

Summaries of Partial or Complete Solutions 89
Directed Line of Reasoning (DLR) 94
Content-specific Explanations 97
Choice of How Interactive to Be 99

VI. CONCLUSION 100

Summary 100
Future Work 101

BIBLIOGRAPHY 108

LIST OF FIGURES

Figure		Page
1.	Grimes' Rhetorical Predicates	21
2.	Williams' Rhetorical Predicates	22
3.	Rhetorical Predicates Defined by McKeown	22
4.	RST-style Diagram For, "You have it slightly, but importantly, wrong."	87

CHAPTER I

INTRODUCTION

Overview and Significant Problems Attacked

I have developed and implemented algorithms for a discourse-generator module that will generate the discourse structure of English text. This English text will be the output of a new version (version 3) of an Intelligent Tutoring System (ITS) called CircSim-Tutor. CircSim-Tutor communicates with the student by generating English text on the screen and accepting English text typed by the student on the keyboard. The tutor productions and the student productions are in almost unconstrained English, and that is an intended feature of the CircSim-Tutor system.

In this thesis, I present studies of student initiatives in keyboard-to-keyboard tutoring dialogue, and the tutors' responses to them. I also explain how we can generate extended summaries, explanations, and directed-line-of-reasoning examples like those produced by our human tutors, Joel Michael and Allen Rovick.

In reviewing the literature, I identified many issues that we can consider in designing and improving CircSim-Tutor. Particularly significant is the suggestion that we may enhance student understanding and recall of material if we try to have CircSim-Tutor take a consistent point-of-view in its explanations and summaries, as well as in its question-and-answer exchanges with the students and the suggestion that CircSim-Tutor should make an obvious point of telling students to reason in terms of physical cause-and-effect, not in teleological terms of purpose. As this is not the usual style of reasoning in biology, CircSim-Tutor should maintain a "high index of suspicion" that students will have a teleological reasoning style.

With respect to the student initiatives in our keyboard-to-keyboard transcripts, I have given a classification scheme for them, described various studies of inter-rater agreement using that scheme, and discussed the results of those studies. In the opinion of all six persons who have used it, the classification scheme does describe the initiatives occurring in a substantial body (about 58 hours worth) of keyboard-to-keyboard tutoring transcripts.

I have developed algorithms to generate summaries, explanations and multi-turn structures such as Directed Line of Reasoning (DLR) exchanges. I have shown that important commonalities underlie all these tutorial discourse structures, particularly with respect to the knowledge that the tutors must consult. Little or no previous work that we are aware of has been done on generating multi-turn structures such as DLRs. I have discovered that in our keyboard-to-keyboard transcripts DLRs serve various roles in tutoring: as summaries, as extended hints, and as a form of explaining a chain of cause and effect. I have also explained how to integrate them into a proposed overall architecture of CircSim-Tutor.

Background of Project

CircSim-Tutor is designed to tutor first-year medical students who are learning how the baroreceptor reflex functions. This reflex compensates for any significant perturbation of the blood pressure. Most of the compensation happens over a period of several seconds, although a steady state may not be reached for a couple of minutes. Given only a textbook description and class lectures, many first-year medical students have difficulty understanding how this reflex works. CircSim-Tutor poses clinical scenarios and asks the student to state, in detail, the events that will occur as the baroreceptor reflex takes effect. As the student solves these clinical problems, the system tutors the student in response to any errors made. If there are no errors, the system probes the student's ability to explain *why* the answers are correct. For the system to do these tasks most effectively, it must be able to generate discourse structure for reasonably complex tutorial dialogues.

When CircSim-Tutor poses clinical scenarios or problems, which perturb the cardiovascular system or blood pressure, and asks the student to state the events that will occur, we break the events in the problem into three phases. First, we tutor about a Direct Response phase before the reflex begins to compensate. Next comes a Reflex Response phase in which we discuss what the reflex will do, and finally a Steady State phase in which we compare the state of affairs before the perturbation to the final steady state result after the reflex compensation.

Explanation, Discourse Generation, and the Subtasks Involved

By the word "discourse," we mean the structure of text at a scale larger than the clause or sentence. The structure of text reflects the structure of the information the text conveys, the intentions or purposes of the writer or speaker, and the focus of attention at each point in the text. Text structure has four important facets. In addition to the purely surface level textual structure, we can speak of text as having informational structure, intentional structure, and attentional structure. Thus, when studying discourse, we are interested in the relationships among parts of the text and the way these relationships are communicated. Some of the meaning of text is conveyed by these relationships directly. Thus, an isolated or unrelated sentence (by itself) is not a discourse; context is the essence of discourse. It is important to note that world knowledge shared by speaker and hearer is part of the context. This context includes common world or cultural knowledge as well as knowledge previously conveyed by the text.

We see Natural Language Generation, of English output, as consisting of two processes. The first process, known as discourse generation, focuses on creating the underlying structure of the text. The second process, surface generation, turns the discourse structure into English sentences. To elaborate, discourse generation involves several decisions: deciding content, deciding what to emphasize (prominent vs. subordinate material), deciding the topic and focus at each point, and deciding the subject matter relationships among the elements of the text.

In the computational linguistics literature, discourse generation is often referred to as "deep generation." Regrettably, for those who do not follow the computational linguistics literature the term "deep generation" suggests that what is generated is the deep structure (DS) of transformational grammar. Using the phrase "discourse generation" will avoid that misunderstanding.

Discourse generation and surface generation must interact, as some surface-level generation choices depend on the discourse structure. For example, the current discourse focus determines what to pronominalize. Let's elaborate on this example. First of all, we need to explain what is meant by the term "discourse focus." At each point in the text, some topic is the current focus that is being talked about. The focus is a part of the attentional structure aspect of discourse structure. Repeated references to the discourse focus are usually indicated by using pronouns in place of the more explicit noun phrase, thus indicating to the reader or hearer that the item being pronominalized is the discourse focus. Thus, surface-level choices of pronominalization depend on the underlying discourse structure choices of focus. Now this means the discourse focus is a facet of text structure, and the focus, in turn, reflects the speaker or writer's current intended subject of attention. The surface-level choice to pronominalize indicates the speaker or writer's intention to sustain the pronominalized subject as the discourse focus.

Obviously, before we can sustain a focus, we must establish that focus. How does something become the current focus? Whenever anything new is mentioned—something that is not already established as a potential focus—it then becomes a potential focus. All the topics introduced into the text for the first time are potential foci. This introductory reference is a descriptive noun phrase and provides some specification that is sufficiently detailed that it will evoke the intended referent in the mind of the hearer or reader. Normally, each such topic is still just a potential focus when we reach the end of the sentence in which it is introduced. An immediate further reference to a potential focus will confirm it as the actual current focus. We then expect all further references to the current focus to be in a less specific form. These less specific forms are typically noun-phrases that are less specific than the initial reference or else pronouns. Continuing references in these less-specific forms indicate that the current focus of the text is unchanged. A common pattern is for the first mention to be a specific detailed noun phrase, the second mention to be a less specific noun phrase, and further references to be pronouns.

When the focus changes, leaving a topic that could plausibly be discussed further, that focus is suspended. Suspended foci show stack-like behavior (behavior resembling a spring-loaded stack of dinner plates in a cafeteria). When an anaphor does not satisfactorily co-refer with the current focus or with the current agent, we expect to find the referent in the stack of suspended foci. The resumptions of a suspended focus typically take the surface form of a pronoun or a rather unspecific noun-phrase. If we do find the referent somewhere in the focus stack, all items that are above the referent on the stack (and thus were added later) are eliminated from future consideration as potential referents of anaphora. To refer to one of these discarded foci subsequently, we would then have to reintroduce it from scratch with a noun phrase, much like an initial reference to anything. We *expect* that foci on the stack may be resumed, and for Computer Science people this *is* a stack-like behavior. Interestingly, many people who have learned about stack data structures in Computer Science or Mathematics courses, notice this stack behavior in their own conversations and refer to it explicitly. It is a particularly noticeable feature of conversations among some software people (hacker-speak).

To summarize discussion of the surface form of an anaphor, we have noted that it depends intimately on the underlying attentional structure of the discourse. In detail, the relevant attentional structure consists of a focus stack, a current focus, a current agent, and possibly the potential foci.

Discourse Generation in the Context of CircSim-Tutor

The task of generating discourse structure for CircSim-Tutor clearly reflects the fact that the text being generated is tutorial dialogue. Any tutoring session consists of alternating tutor and student turns. Tutorial dialogue discourse structure can be divided into pieces at multiple scales. One aspect of this is that the resulting size of text produced for a single turn varies greatly. Another aspect is that a tutor has some goals that can be satisfied in a single turn but other goals that require several turns. I will now discuss these issues of scale in more detail, providing examples that I consider important for my research.

Consider the amount of text generated for any single turn. A turn may be quite brief or may be substantial. For example, the tutor can make a single sentence statement, ask a single sentence question, or give an acknowledgement such as “correct” or “no.” At other times, the tutor will produce an extended paragraph. Extended paragraphs are normal when the tutor gives detailed explanations or summaries of the correct reasoning for the material discussed in the preceding several turns. On more rare occasions, we find tutors producing a turn with multiple paragraphs. This typically seems to occur with a weaker student for whom a tutor eventually resorts to giving a small didactic lecture.

A different sort of larger scale discourse structure is required for multi-turn structure. We find that human tutors often have goals in tutoring that require multiple turns to accomplish. One important example of this is when a tutor leads a student step by step through a line of reasoning that the student is having problems with. Usually, this is done with a series of *bite-size* questions. We refer to this as a “directed line of reasoning” or DLR. An open question for us is whether there are other kinds of multi-turn structure that are similarly clear cut. There are many obvious ways in which multi-turn structure differs from text produced by a single speaker or writer. A particularly important difference is that the student’s turns are not completely predictable. In a directed line of reasoning, the student may give the expected answer, but instead may give wrong answers, and may ask questions or otherwise take the initiative. Thus, replanning may be required.

Instances of discourse structure on a smaller scale that are important for CircSim-Tutor include positive and negative acknowledgments of statements by the student, hints, and simple questions that CircSim-Tutor asks the student. Considerations of topic and focus, references to previous material, and references to future material, are involved in integrating these productions into the tutorial dialogue. In producing these smaller scale structures, the discourse generator mainly handles the topic and focus aspects. Because sustaining topic and focus are discourse issues, there is an important larger scale model that the discourse generator must maintain, even when the tutor productions are on a smaller scale. This *requires* the discourse generator to participate in generating these smaller scale productions.

Although my work will focus to some degree on larger-scale structures, CircSim-Tutor still has serious problems with the smaller scale structures. The poor quality of the hints in version 2 of CircSim-Tutor is a particularly severe problem. I feel the solution to these problems is primarily a task for modules other than the discourse generator. As is true for most of the smaller scale structures that present problems, the important improvements in hinting should come from improved planning of tutoring tactics (in the instructional planner).

Contributions of this Work

An important part of this thesis is work on the larger scale or multi-turn structure of tutorial dialogue. Two parts of this work on larger scale discourse structures appear to be significant results. First, little or no work that we know of has been done on multi-turn structures in Intelligent Tutoring Systems. In this thesis I have given algorithms that I developed to generate directed-line-of-reasoning (DLR) structures. I have shown how they can be integrated into the architecture of the CircSim-Tutor system, and I have shown multiple roles (summaries, extended hints, and a form of explaining a chain of cause and effect) that DLRs play in our transcripts of keyboard-to-keyboard tutoring sessions. Second, I have shown that important commonalities underlie the generation of summaries, explanations, hints, and DLRs, particularly with respect to the knowledge that tutors must consult in generating them.

Usually the tutor has the initiative in our tutoring sessions. By “student initiatives” we mean instances where the student is trying to alter the course of the session, for example by asking a question. One important aspect of the discourse structure in a tutoring dialogue is exchanges triggered by student initiatives. These exchanges occur when the student asks a question or otherwise takes the initiative in the session and the tutor responds. I have done an extensive analysis of about 58 hours worth of keyboard-to-keyboard transcripts in which our human tutors tutored first-year medical students on the same material that CircSim-Tutor teaches. The main focus of that analysis was to determine how students take the initiative and how we can make CircSim-Tutor recognize occurrences of student initiative. In addition, I have identified several factors that appear to influence how our human tutors responded to the initiatives in our transcripts. Further work will be needed on how to respond when a student initiative is recognized.

The Text Generation vs. Cognitive-Psychology Senses of the Word "Schema"

When used in the sense of "text generation from schemas," the word "schema" refers to a sort of template for a piece of text. The general idea is most easily grasped by seeing several examples, and several are given in Chapter II of this thesis. But, this text generation sense of the word "schema" must be distinguished from the (more familiar) sense commonly encountered in cognitive psychology, where the word really refers to a kind of meta-construct that organizes related knowledge, beliefs, or attitudes. Both senses of the word are of necessity used in this thesis. To help the reader keep these two senses straight, I take a brief detour here to explain the cognitive psychology sense of the word.

In the cognitive psychology sense (see, e.g., the chapter by Winfrey and Goldfried [1986]), a schema is a prototypical abstraction of a complex concept, and includes examples as well as relationships among sub-elements. In addition, schemas combine structure and processing (learned mental behavior routines). Schemas are induced from repeated past experience with many exemplars of the concept. They appear to have some key characteristics: to be organized hierarchically, to increase efficiency of processing (especially retrieval of information or of emotional attitudes), to have their original motivation or source “compiled out”, and correspondingly to be very robust so that information conflicting with an existing schema is processed with difficulty and has a strong tendency to be rejected (in fact, schemas tend to be modified to have special cases rather than to be rejected entirely, so that getting rid of a truly mistaken schema may require a carefully directed frontal assault).

One can easily find schemas (or sets of schemas) for which the person holding the schemas has great trouble explaining their genesis. Correspondingly, in the case of novices one can expect to find significant confabulation of justifications for existing schemas. The almost unconscious nature of schemas is actually a central part of the psychological theory of schemas. Extracting schemas is a classical knowledge engineering task. Conveniently for us, analyzing the students’ cognitive processes in terms of schemas does not require us to determine the genesis or purpose of the schema.

To give examples of schemas, one could speak of a “game” schema, an “implausible animals” schema, a “pretty music” schema, and so forth. We hypothesize a set of “tutoring the baroreceptor reflex” schemas. More to the point, one can speak of a circulatory-system schema, a systemic-circulation schema, a venous-system schema, and so forth. In fact, some crucial student misconceptions appear to me to stem from faulty schemas in this sense—for example the error where a student says that increased cardiac output is immediately reflected in increased venous return thus causing an immediate increase in central venous pressure or right atrial pressure. In this error, the student apparently has a schema of the systemic circulation as sort of a short non-compliant tube.

The robust persistence of schemas also appears to me to account nicely for the observed tendency of Rush Medical College students to disbelieve what they are being taught about the baroreceptor reflex and associated circulatory physiology. A great many examples of this occur in the transcripts, and we feel the most striking examples turned up in a study that we did where the tutoring was done by inexperienced student tutors. These tutors were trained relatively intensively. These student tutors are medical students who had been through the physiology class, had run CircSim until proficient with it, and they demonstrably knew the material well enough to be proficient at solving the clinical problems that they were tutoring. In the transcripts of the sessions where these student tutors served as the tutor, one finds examples where they offer incorrect justifications for how the physiology works and one finds instances where they offer correct justifications and then actually say, "Dr. Michael and Dr. Rovick say that <*some correct physiology statement*> but I don't really believe it.") We have no alternative explanation for why these medical students would disbelieve their physiology professors.

CHAPTER II

REVIEW OF THE LITERATURE

Work on rhetoric and effective communication dates from antiquity. Naturally, work on computer generation of discourse and dialogue is more recent. In this review, I restrict myself to computer generation of text, first treating discourse generation in general and then treating several recognized subproblems.

Text Generation as Choice or Planning

The prevailing view is that text generation is based on making choices and perhaps on planning the text in advance. Douglas Appelt [1985a, 1985b] describes text generation as a process of hierarchical nonlinear planning. He exploits the planning formalism called a "procedural network" that he states was used in Earl Sacerdoti's NOAH planner. The underlying idea of that system is to plan individual goals separately, then modify and combine the plans using "critic" procedures. This approach to planning creates a partial ordering of the individual goals and, at least in theory, provides opportunities for parallel processing during the phase when individual goals are being planned separately. Since the planning process is based on goal satisfaction, it follows that some procedure(s) must monitor the planning process and notice that we are done when all the goals have been satisfied and we are done merging their solutions. Of course, this task of deciding when we are done and can quit is really an instance of the Frame Problem [McCarthy and Hayes, 1969].

Appelt exploits the critic routines to find opportunities to satisfy multiple goals in one sentence. For example, referring expressions—noun-phrases (NPs) or physical gestures which refer to physical objects—may introduce objects to the hearer's or reader's attention, in addition to playing some other role in a sentence. Now, this is a very interesting accomplishment for a text planner, since human use of language also does this with referring expressions. Later text planners written by other researchers have not usually included this ability. Some specific examples of multiple goals are referring plus communicating additional information, referring plus communicating an emotional attitude toward the object referred to, or requesting plus being polite.

Interesting aspects of the implementation of Appelt's work are its uniformly planning-based approach and the lack of separation between deciding what to say and deciding how to say it. Most other planners decide what to say first and only later decide how to say it. As a result, these other planners have fewer opportunities for interaction between the two tasks (what to say and how to say it).

There is an obvious "speech acts" flavor to Appelt's work. He observes that the process of understanding language entails that the hearer must interpret the speaker's illocutionary intentions. James Allen [1983b] provides some interesting examples: "Could you tell me the time?" "The train to Montreal?" where a hearer who is trying to be maximally helpful must infer the speaker's plan as well as the speaker's intent. On the other hand, intended perlocutionary effects are not necessarily intended to be recognized, even though they are intentions behind the speaker's planning.

The grammar (called TELEGRAM) in Appelt's system is unification based. It starts with a minimal description of the speech act and calls the planner whenever the unifier needs more information. As a result, descriptors in the text get bound as late as possible, after more information about the entire sentence being planned is available. This appears to me to be a key advance because the choice of descriptors occurs only after as much information as possible is available.

On the other hand, Appelt's system is really generating a sentence at a time. One can argue that this is a weakness, despite the system's comprehensive approach to planning each sentence and despite the ability to generate sentences that satisfy multiple goals. Because it does not plan discourse on a larger scale, I do not believe that Appelt's approach can be made to work well as the discourse and dialogue planner for a system like CircSim-Tutor.

David McDonald [1983], working a couple of years before Appelt, found modern theories of linguistic competence to be inadequate for goal-directed text generation, stating that, "their formal structure does not permit them to address the central problem, i.e., *how specific utterances arise from specific communicative goals in a specific discourse context.*" [McDonald, 1983, pg. 209] So McDonald began with questions about what information and representation natural language generation starts with, what kinds of decisions are made, how are they controlled, and what sorts of intermediate representations are needed. Like Appelt, McDonald based his system on planning, but unlike Appelt, McDonald kept a sharp distinction between a "speaker program" and a "linguistic component." As McDonald described his system, the speaker decides content and does any extensive text planning that is required. The linguistic component is fast and does minimal planning. The interface between them is a "message," and it is clear McDonald regards the process of producing language as a translation into English from something like mentalese or from a message produced by the speaker program. McDonald has used the same linguistic component with six different speaker programs. Viewed as planning, McDonald's program has the internal goal of translating a message into English.

The linguistic component uses a dictionary and a grammar. What is looked up in the dictionary is really messages: it is not an English lexicon. The dictionary is specific to each speaker program. The primary idea of the linguistic component is to create a tree-structured message and to rewrite the tree, in a basically recursive-descent fashion. McDonald points out that the well-defined nature of the interface between speaker and linguistic components implies, "that the extent of their shared assumptions about representations and contingencies is small (otherwise the linguistic component would have to be largely rewritten for each new speaker)" [McDonald, 1983, pg. 211].

The paper describes in detail the generation of English text from a set of predicate calculus sentences which constitute a "barber proof." The generated text for the proof is as follows

> Assume that there is some barber who shaves everyone who doesn't shave himself (and no one else). Call him Giuseppe. Now, anyone who doesn't shave himself would be shaved by Giuseppe. This would include Giuseppe himself. That is, he would shave himself, if and only if he did not shave himself, which is a contradiction. This means that the assumption leads to a contradiction. Therefore, it is false, there is no such barber. [McDonald, 1983, pg. 216]

In discussing the barber proof given above, McDonald [1983] pointed out the following significant accomplishments embodied in the current version of his generation program. (1) The ability to go beyond the literal content of the message produced by the speaker (predicate calculus). (2) Subsequent reference, with correct handling of pronominalization and non-pronominalization, including some instances of pronominalizing descriptions ("no such barber"). (3) Referring to predicate calculus sentences in the proof by functional labels, such as "assumption" and "a contradiction." (4) Context-sensitive realization. (5) Attempts to avoid ambiguity, such as the parenthesized phrase in the generated barber proof.

McDonald also provides detailed discussion of generation from a set of PLANNER-style assertions. He identified the following significant accomplishments of his program in this domain. (1) Varying the paragraph structure. (2) Omitting "given" information. (3) Varying descriptions with context, particularly varying the level of detail appropriately. (4) Using ellipsis.

McDonald imposed the following four restrictions on his model: (1) on-line operation (both parts of the linguistic component must process the current message element before the speaker produces the next element), (2) indelibility (modeling a speaker rather than a writer), (3) locality of context in making decisions (no scanning the tree for information), and (4) real-time operation (bounded processing).

McDonald regards his program as evidence for the hypothesis that if the program succeeds and fails in the same ways as spoken human language and if it has the computational characteristics described in the previous paragraph (with its computational characteristics necessarily falling out from the algorithms and invariants of the data structures) then this suggests the program is isomorphic to the processes involved in human production of language. I would characterize this as a claim of *psychological validity*, although it is obviously also a claim about the computational complexity of human language.

McDonald describes the linguistic component as being appropriate when discourse tasks are required (embedded clauses, coherence relations in multi-sentence texts, context sensitive realizations, and describing objects from their properties rather than having canned descriptions). McDonald's program cannot do the following: creative expression such as fitting old words to new situations, monitoring itself (as a planning technique), recognizing when a message will unavoidably lead to awkward or ungrammatical text, reasoning about trade-offs caused by limited expressibility, and planning by backwards chaining from desired linguistic effects.

Text Generation from Schemas

When used in the sense of "text generation from schemas," the word "schema" refers to a sort of template for a piece of text. Text generation from schemas appears to have been developed by Kathleen McKeown [1985]. The general idea is most easily grasped by seeing several examples, and several are given below. But, this text generation sense of the word "schema" must be distinguished from the (more familiar) sense commonly encountered in cognitive psychology, where the word really refers to sort of a meta-construct that organizes related knowledge, beliefs, or attitudes. McKeown borrowed the word "schema" from cognitive psychology and used it to name the framework she had developed. Both senses of the word are of necessity used in this thesis. The cognitive psychology sense of "schema," which was described at the end of Chapter I of this thesis, is *not* what is being used in the remainder of this section. While reading the following discussion, the reader may perhaps form a schema (in the cognitive psychology sense) for text-generation schemas.

McKeown [1985] built an influential text-generation system around schemas. McKeown's system, called Text, provides information about the structure and contents of the Office of Naval Research (ONR) database that contains information about ships and destructive devices. Specifically, Text performs three tasks: define an item, describe an item, and differentiate/compare two items. Examples of questions that Text can answer include:

- What is a frigate?
- What do you know about submarines?

- What is the difference between an ocean escort and a cruiser?

McKeown's system uses schemas, composed of rhetorical predicates such as those described by Joseph Grimes, which McKeown describes. McKeown found it necessary to invent three new kinds of predicates: identification (proper noun, etc.), renaming (A.K.A.), and positing. McKeown used four kinds of schemas: identification, constituency, attributive, and comparison. Each is described in more detail below. I note that schemas, although a static representation, worked well for the rather static tasks that Text performs.

One can take a skeptical, even critical, attitude toward schemas. Schemas are really only structural or rhetorical. They do not seem to me to account for intentional or attentional structure, as described by Grosz and Sidner [1986]. Moore and Pollack [1992] discussed this issue, reaching the same conclusions. Schemas are too inflexible for dialogue in my opinion. Moore and Paris made similar observations, stating

> Schemata (McKeown, 1985) encode standard patterns of discourse structure, but do not include knowledge of how the various parts of a schema relate to one another or what their intended effect on the hearer is. A schema can be viewed as a compiled version of [their text plans] in which all [the planning information] has been stripped out and only the leaves (the speech acts) remain. While schemata can produce the same initial behavior as one of our text plans, all of the *rationale* for that behavior has been compiled out. Thus schemata cannot be used to participate in dialogues. If the user indicates that he has not understood the explanation, the system cannot know which part of the schema failed to achieve its effect on the hearer or which rhetorical strategy failed to achieve this effect. [Using the Text system] is essentially planning a schema from more fine-grained plan operators. . . . improving the flexibility of the system. [Moore and Paris, 1989, pg. 209]

McKeown gives typical content for the four kinds of schemas. The Text system actually implements simplified versions of all four. The content consists of various rhetorical predicates. Rhetorical predicates are generally a single sentence or clause that performs a specific rhetorical function in a text. McKeown used rhetorical predicates that had been identified by Joseph Grimes and W. Williams, plus some additional predicates that she found useful. McKeown lists the rhetorical predicates, with examples of each, that she took from each source. See the following three figures (Figs. 1–3), taken from McKeown [1985], for details.

McKeown gives the structure of her schemas by using a formal syntactic notation. In the examples following Figure 3, curly braces indicate optional elements. Superscript '*' and '+' are the Kleene operators. The '|' character indicates alternatives. A ';' indicates ambiguity. I have inserted parentheses to indicate groupings.

For the convenience of readers who may not be familiar with this sort of syntactic formalism, let me provide a few simpler examples to illustrate the meaning of all the symbols. I will underline the examples, to set them off visually. Anything followed by a superscript '+' must occur one or more times. Thus, $\underline{X^+}$ indicates a string of at least one X and possibly a great many X's. A superscript '*' is the same as a superscript '+' except that it means zero or more occurrences rather than one or more. Thus anything followed by a superscript '*' can be considered optional. In fact, $\underline{X^+}$ is equivalent to $\underline{XX^*}$, and both really mean an X followed by zero-or-more additional X's. The curly brace symbols indicate an optional constituent. As the reader might expect, this implies that $\underline{\{X^+\}}$ is equivalent to $\underline{X^*}$. The vertical bar '|' symbol indicates alternatives. Consider the example $\underline{(X|Y|Z)^+}$, which indicates a string of at least one letter (one or more) and that the available letters (the alternatives) are X, Y, and Z. The parentheses in that example indicate the scope of effect of the superscript '+', so that it applies to the three alternatives as a group rather than to just the Z. Thus the following are all strings that match $\underline{(X|Y|Z)^+}$. X Y Z XX XY YX YY YZ ZY ZZ XXX XXY XYX YXX

XYY YXY YYX YYY, etc. Reference to these examples should enable the reader to understand the following specifications of the four types of schemas defined by McKeown. Notice that each of the four schemas includes, as a required part, a rhetorical predicate having the same basic purpose as the schema as a whole.

McKeown's identification schema identifies an entity or event.

Identification Schema [McKeown, 1985, pg. 28]
Identification (class and attribute)
{Analogy | Constituency | Attributive | Renaming | Amplification}*
(Particular-illustration | Evidence)+
{Amplification | Analogy | Attributive}
{Particular-illustration | Evidence}

Minimally, the identification schema thus consists of:
Identification (class and attribute)
Particular-illustration | Evidence

The following paragraph can be analyzed using McKeown's identification schema.

> Eltville, Germany, is an important wine village of the Rheingau region (*identification—class and attribute*). The vineyards make wines that are emphatically of the Rheingau style (*attributive*), with a considerable weight for a white wine (*amplification*). Taubenberg, Sonnenberg, and Langenstuck are among vineyards of note (*particular illustration*).

- Attributive
  Mary has a pink coat.

- Equivalent
  Wines described as "great" are fine wines from an especially good village.

- Specification (of general fact)
  Mary is quite heavy. She weighs 200 pounds.

- Explanation (reasoning behind an inference drawn)
  So people form a low self-image of themselves, because their lives can never match the way Americans live on the screen.

- Evidence (for a given fact)
  The audience recognized the difference. They laughed from the first frames of that film.

- Analogy
  You make it in exactly the same way as red-wine sangria, except that you use white wine.

- Representative (item representative of a set)
  What does a giraffe have that's special? — a long neck.

- Constituency (presentation of sub-parts or sub-classes)
  This is an octopus . . . There is his eye, these are his legs, and he has these suction cups.

- Covariance (antecedent, consequent statement)
  If John went to the movies, then he can tell us what happened.

- Alternatives
  We can visit the Empire State Building or call it a day.

- Cause-effect
  Adding spirit during the period of fermentation arrests the fermentation development.

- Adversative
  It was a case of sink or swim.

- Inference
  So people form a low self-image of themselves.

Figure 1. Grimes' Rhetorical Predicates [McKeown, 1985, pg. 23]

- Comparison
- General illustration
- Amplification
- Conclusion
- Topic
- Particular illustration
- Contrasting

> *What, then, are the proper encouragements of genius?* (topic) *I answer, subsistence and respect, for these are rewards congenial to nature.* (amplification) *Every animal has an aliment suited to its constitution.* (general illustration) *The heavy ox seeks nourishment from earth; the light chameleon has been supposed to exist on air.* (particular illustration) *A sparer diet than even this satisfies the man of true genius, for he makes a luxurious banquet upon empty applause.* (comparison) *It is this alone which has inspired all that ever was truly great and noble among us. It is as Cicero finely calls it, the echo of virtue.* (amplification) *Avarice is the pain of inferior natures; money the pay of the common herd.* (constrasting sentences) *The author who draws his quill merely to take a purse no more deserves success than he who presents a pistol.* (conclusion)

Figure 2. Williams' Rhetorical Predicates [McKeown, 1985, pg. 24]

- Identification
  ELTVILLE (Germany), an important wine village of the Rheingau region.

- Renaming
  Also known as the Red Baron.

- Positing
  Just think of Marcus Welby.

Figure 3. Rhetorical Predicates Defined by McKeown [McKeown, 1985, pg. 25]

McKeown's constituency schema gives the parts and subparts of an entity or event.

<u>Constituency Schema</u> [McKeown, 1985, pg. 29]

Constituency

Cause-effect* | Attributive* |

{ Depth-identification | Depth-attributive

{Particular-illustration | Evidence }

{Comparison | Analogy} **}**+

{Amplification | Explanation | Attributive | Analogy}

Minimally, the constituency schema thus consists of:
Constituency

The following paragraph can be analyzed using McKeown's constituency schema.

> Steam and electric torpedoes: Modern torpedoes are of two general types (*constituency*). Steam-propelled models have speeds of 27 to 45 knots and ranges of 4000 to 25,000 yards (*depth-identification/depth-attributive*). The electric powered models are similar (*comparison*), but do not leave the telltale wake created by the exhaust of a steam torpedo (*amplification*).

McKeown's attributive schema can illustrate a point about some concept or object.

Attributive Schema [McKeown, 1985, pg. 27]
Attributive
{Amplification ; Restriction}
Particular-illustration*
{Representative}
{Question ; Problem
Answer} |
{Comparison ; Contrast
Adversative}
Amplification | Explanation | Inference | Comparison

Minimally, the attributive schema thus consists of:
Attributive
Amplification | Explanation | Inference | Comparison

The following paragraph can be analyzed using McKeown's attributive schema.

> This book, being about work, is by its very nature about violence (*attributive*) — to the spirit as well as to the body (*amplification*). It is about ulcers as well as accidents, about shouting matches as well as fistfights, about nervous breakdowns as well as kicking the dog around (*particular illustration*). It is, above all (or beneath all), about daily humiliations (*representative*). To survive the day is triumph enough for the walking wounded among the great many of us (*amplification; explanation*).

Compare and Contrast Schema [McKeown, 1985, pg. 31]
Here, A is the main point and ~A is the negative point, ¬ A (i.e., *not* A)

Posiing | Attributive (~A)
{Attributive (A) |
Particular-illustration | Evidence (A) |
Amplification (A) |
Inference (A) |
Explanation (A) }$^{+}$
{Comparison (A and ~A) |
Explanation (A and ~A) |
Generalization (A and ~A) |
Inference (A and ~A) }$^{+}$

More abstractly, the compare and contrast schema consists of:
stuff on (~A)
stuff on (A)
comparison or contrast of (A and ~A)

The following paragraph can be analyzed using McKeown's compare and contrast schema.

> Movies set up these glamorized occupations (*positing: ~A*). When people find they are waitresses, they feel degraded (*attributive: A*). No kid says I want to be a waiter, I want to run a cleaning establishment (*evidence: A*). There is a tendency in movies to degrade people if they don't have white-collar professions (*comparison; explanation: A and ~A*). So, people form a low self-image of themselves (*inference: A and ~A*), because their lives can never match the way Americans live—on the screen (*comparison; explanation: A and ~A*).

McKeown used an Augmented Transition Network (ATN) in generation, selecting and adding a "proposition" each time a "fill arc" was traversed. Additionally, as is typical of ATNs, McKeown has a set of registers. Her registers are as follows: current focus, global focus, potential focus list, knowledge pool, question type, and the currently built part of the output message. At each state of the ATN, all possible next states are computed, and focus constraints are used to choose the actual next state.

McKeown's algorithm for selecting focus, which she gives in a fair amount of detail, is based on a set of focus constraints:

- Shift to an item mentioned in the previous sentence
- Maintain current focus
- Return to a topic of earlier discussion
- Select proposition with the greatest number of links to the previous sentence.

Use of these focus constraints greatly increases the coherence of the text she generates.

The Text system uses a fairly simple approach to partitioning off the relevant knowledge in the knowledge base. The approach to picking relevant knowledge for comparisons is fairly complex, but McKeown states that doing this was important to the performance of the system. We should consider the techniques she used. There is no model of the specific user. The grammar is based on Martin Kay's Functional Grammar. Text was implemented in CMU Lisp on a VAX 11/780. To answer a typical question, the (uncompiled) "strategic component" (deep generation) took about 1½ minutes, and the (compiled) "tactical component" (surface generation) took about 20 minutes.

Earlier thinking in our project [Zhang, 1991] considered schemas as a basis for text generation in CircSim-Tutor. Yuemei Zhang felt that schemas, based on the ideas of McKeown, would capture the structure of our human tutors' contributions to tutoring sessions. Zhang seemed to me to believe that one could enumerate the recurring tutor contributions, thus allowing us to build a schema for each. In addition, her schemas were hierarchical and as a result they did capture some very important generalizations and abstractions about tutoring. However, for reasons covered in the discussion of Kathleen McKeown's work (and that of Moore and Paris), I do not think a schema-based approach can work well for generating tutorial dialogue in CircSim-Tutor. For this reason, I have not chosen to follow up directly on Zhang's discourse organization, despite the interesting character of her analysis.

Anaphoric References

We turn now to work done by Bonnie Lynn Webber [1983], in which she explained what entities are available for anaphoric reference at any given point in a text. Webber's theory is based on the structure, rather than content, of the text. The important contribution of Webber's work is that it gives an implementable account for inter-sentential pronoun reference and other inter-sentential anaphora. Webber states that her theory is based on the idea of a "discourse model" that is synthesized by the hearer or reader. Webber explains that by "discourse model" she means the items "naturally evoked"—mostly by noun phrases (NPs)—and the relationships in which they participate.

To some degree, what entities are evoked depends on the surface form of the sentence(s): the internal "logic form" representation is not sufficient to account for the anaphoric referents. *Combinatoric* features, such as interaction, dependency, and cardinality are also important in resolving anaphora. Discussing the role of discourse focus in anaphora, Webber observes,

> A speaker is usually not able to communicate all at once the relevant properties and relations s/he may want to ascribe to the referent of a discourse entity. To do that, s/he may have to direct the listener's attention to that referent (via its corresponding discourse entity) several times in succession. When the speaker wants to re-access an entity already in his/her [discourse model] (or another one directly inferable from it), s/he may do so with a definite anaphor (pronoun or NP). . . . The problem then, at least for definite anaphora, is

> identifying what discourse entities a text naturally evokes. [Webber, 1983, pp. 336–337]

Webber discusses definite pronoun anaphora and *one* anaphora. She lists the following features of sentences, which her theory has to handle [Webber, 1983, pg. 339]: (1) distinguishing between definite and indefinite noun phrases and between singular and plural noun phrases; (2) distinguishing, for each modifier in a plural noun phrase, whether it conveys information about the entire set denoted by the plural noun phrase or about the individual set members; (3) resolving any ellipsed verb phrases in the sentence; and (4) identifying what have traditionally been called "quantifier scope assignments," although . . . they may not be determinable when the sentence is first heard.

The importance of these distinctions is that they determine what entities are evoked (via their descriptions) and thus are available to talk about or refer to. The hearer must synthesize descriptions for indefinite noun phrases, taking the context into account. The hearer also needs to recover ellipsed verb phrases (VPs), as referents of anaphors may be in an ellipsed VP. Webber gives the following example. "John didn't bake a cake for Wendy. On the other hand, Elliot did, but she didn't like it." [Webber, 1983, pg. 341] To deal with these phenomena in CircSim-Tutor we, too, would need to recover ellipsed verb phrases, and so forth.

Webber proposes, in detail, a quantified internal representation to handle all these issues. Webber wants to handle each of the combinatoric features independently, because many people have trouble with some quantified sentences. Separating their representations would allow separating the aspects people will definitely understand from the aspects they may not. In handling plural noun phrases, Webber distinguishes various types of plural readings that correspond to different quantifier scope assignments. Webber also discusses the handling of generic sets, as seen in the following example. “Joe has a Japanese car. They’re selling quite well.”

I believe this work has two implications for a discourse or dialogue generator. First, it supports the view that the discourse generator must *tell* the surface generator when some item referred to in the discourse is going to become a focus. Second, the discourse generator must tell the surface generator about quantifiers. At this point, other members of the CircSim-Tutor development team want to deal with quantifiers in much less detail than in the internal representation that Webber created. So, I plan to omit many of the instances of quantifiers that Webber includes.

Having a Point of View Enhances Understanding and Recall

Robert P. Abelson [1975b] published the results of a study that I consider quite important. Abelson begins with the observation, “People with different cognitive styles can become quite exercised over whether the propositional or simulational account is the ‘correct’ psychological description. . . .” [Abelson, 1975b, pg. 140] He then offers evidence that even if some persons answer many questions using propositional knowledge, they resort to mental or physical simulation when they do not know an answer. Abelson’s experiment was about mental simulation of leaving a hotel and strolling a block down a street while hearing a 68-sentence story about a character leaving a hotel and strolling a block down a street. The experiment then studied recall of the story.

Subjects were divided into three groups given different instructions. Then the story was read to them. The first group, the “self group,” was instructed to imagine themselves to be the main character in the story. The second group, the “balcony group,” was told to imagine themselves seeing the story from a fourth-floor balcony. The third group was given no instruction about point of view. The story was read to the three groups in common.

The "self" group had better recall of details involving body sensation, such as aggravating a sore arm or drinking hot coffee. The balcony group had better recall of far visual details such as a sign over a bank. The third group had, relatively, poor free recall of all details. Most of this relative difference disappeared on cued recall.

Near visual details, such as reading a wristwatch, proved more complex. Abelson notes

> In a pilot study, we found that Balcony subjects sometimes reported "floating down" off the balcony from time to time, as it were, to peer vicariously over the actor's shoulder at details otherwise too small to imagine seeing clearly from a distance. . . . the experiment shows different specialized processing modes are keyed to different vantage points, and have different consequences for what is best remembered. Furthermore, the No Vantage Point group seems to suffer in free recall from the lack of a special processing mode. Perhaps the vantage point provides a set of higher-order nodes in the network representing the story, facilitating access to the lower-order story details. [Abelson, 1975b, pg. 142?]

I consider this paper to have interesting similarities to an experiment by Fass and Schumacher [1981]. They showed that an "encoding perspective," such as Abelson describes will have significant effects on what can be retrieved on delayed recall, suggesting ties to schema theory and the preferential learning and recall of the information important in the schema.

Fass and Schumacher had subjects read a passage containing information important to burglars and homebuyers. Subjects were given either a burglar, home-buyer, or no (encoding) perspective during reading, and either a burglar or home-buyer (retrieval) perspective during recall. Subjects were required to recall the passage either immediately after reading or else after a 24 hour delay. The results [Fass and Schumacher, 1981, pp. 21–22] indicated that during immediate recall, either an encoding or retrieval perspective influenced recall. However, when recall was delayed, only the encoding perspective influenced recall. Fass and Schumacher interpreted the results in terms of cognitive psychology schema theory and the encoding specificity principle described below. Schemas, the reader will recall, are frame-like structures hypothesized as the organization of memory. The idea of *encoding specificity* is that having the same retrieval perspective as encoding perspective should improve recall. This was *not* borne out. The paper ends by stating, “It is apparent, . . ., that long-term retention necessitates that the information be organized into a coherent integrated unit. Unless this occurs, the information will not be recalled as well” [Fass and Schumacher, 1981, pg. 25]. I believe this experiment supports our goals of adding discourse organization and a consistent point of view to the language generation of CircSim-Tutor.

Speech Acts

in the William James lectures that he delivered at Harvard in 1955, J. L. Austin [1975] laid the foundation for later theories of speech acts. Austin distinguished constative utterances (roughly, having to do with states of affairs rather than acts) from performative ones, and he analyzed performative utterances in considerable detail. He divided performatives into five general classes: verdictives, exercitives, commissives, behabitives, and expositives [Austin, 1975, pp. 151–164]. Drawing something of a distinction between performatives and descriptives, Austin put the behabitive class somewhere in the middle [Austin, 1975, pp. 83–90]. Interestingly, Austin distinguished direct from indirect performatives [Austin, 1975, pg. 32] What Austin is most recognized for is the definition of the *locutionary* act, *illocutionary* act, and *perlocutionary* act involved in an utterance. These are the actual act of speaking, the intention of the act of speaking, and the side-effects that result from the act of speaking, respectively. Speech act theory is built mostly around the illocutionary component of an act of speaking.

John Searle [1969, 1975] introduced the term "speech act" and examined the conditions under which an utterance counts as a speech act. In [Searle, 1975] he extended his original work to apply to indirect speech acts. Searle stated

> The reason for concentrating on the study of speech acts is simply this: all linguistic communication involves linguistic acts. The unit of linguistic communication is not, as has generally been supposed, the symbol, word, or sentence, or even the token of the symbol, word, or sentence, but rather the production or issuance of the symbol, word, or sentence in the performance of the speech act. . . . More precisely, the production or issuance of a sentence token under certain conditions is a speech act, and speech acts . . . are the basic or minimal units of linguistic communication. [Searle, 1969, pg. 16]

He goes on to say that a speech act must by definition result from the intention that it be an act of linguistic communication. "Only certain kinds of intention are adequate for the behavior I am calling speech acts." [Searle, 1969, pg. 17] Searle extended Austin's distinction among illocutionary acts and perlocutionary acts, stating,

> (a) Uttering words (morphemes, sentences) = performing *utterance acts.*
> (b) Referring and predicating = performing *propositional acts.*
> (c) Stating, questioning, commanding, promising, etc. = performing *illocutionary acts.* [Searle, 1969, pg. 24]

And he goes on to note that one cannot do (a) and/or (b) without doing (c). The basic intentional conditions that Searle identifies are as follows: the propositional content of the speech act, preparatory conditions that would enable the act indicated by the speech act, sincerity on the part of the participants in the speech act itself, and that the essence of the identified speech act actually holds. [Searle, 1969, pp. 57–71]

James McCawley [1988] discussed participant roles in speech acts. He pointed out that there are three possible roles that the speaker may play in a speech act: the author, the utterer, and the actor. These roles may be filled by one to three persons, and this is syntactically marked. McCawley gave a similar analysis for the hearer. He also discussed paragraph grammar, making the interesting observation that a sentence summarizing a preceding stretch of discourse cannot be conjoined with surrounding sentences.

Ceteris Paribus Rules Governing Normal Conversation

H. Paul Grice [1975] proposed a set of conversational rules, which are thoroughly accepted as accurate for normal conversation. Interestingly, however, they may not hold for tutorial dialogue. Grice proposed his set of "conversational maxims" as a result of studying the process of conversational implicature. The maxims really fall out from the observation that at each point in a conversation,

> SOME possible conversational moves would be excluded as conversationally unsuitable. We might then formulate a rough general principle . . . namely: Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. One might label this the Cooperative Principle [Grice, 1975, pg. 45].

Grice makes the point that the mechanism of conversational implicatures involves violating one or more of his maxims, in a way that is recognized by the hearer. He introduces his maxims as being a more specific statement the Cooperative Principle. Grice's maxims are as follows [Grice, 1975, pp. 45–47].

Quantity

- Make your contribution as informative as is required (for the current purposes of the exchange).
- Do not make your contribution more informative than is required.

Quality — Try to make your contribution one that is true.

- Do not say what you believe to be false.
- Do not say that for which you lack adequate evidence.

Relation

- Be relevant.

Manner — Be perspicuous.

- Avoid obscurity of expression
- Avoid ambiguity.
- Be brief (avoid unnecessary prolixity).
- Be orderly

"And one might need others."

Grice states that he is echoing Kant in calling his categories Quantity, Quality, Relation, and Manner.

We have been thinking about, and discussing, how tutors do or do not follow Grice's maxims. Interestingly, tutors appear to violate these maxims intentionally. We might state a general maxim that tutors should not tell students anything that the students might plausibly be able to figure out for themselves. Tutors often fail to answer questions or provide relevant information as a result of this tutoring maxim. Further, tutors may well intentionally give students a false or irrelevant statement in the process of exploring the students' abilities to distinguish true and relevant material. We consider this to be a possible area for further research. Our current intention, in the design of CircSim-Tutor, is simply to have CircSim-Tutor hold off on providing requested or relevant information if we believe the student already knows the information and might well produce it if given a hint or asked an appropriate leading question that would direct the student's attention to the relevant information. In this regard, the interested reader may wish to read the more detailed coverage of this topic in the work of Gregory D. Hume et al. [Hume, Michael, Rovick, and Evens 1993, 1995].

Focus and Its Implications

In this section, I discuss work by Barbara Grosz, Candace Sidner, and Rachel Reichman. In a paper drawing on her 1977 PhD thesis, Grosz pointed out [1977] that domain knowledge is overwhelmingly large and that focus will usefully constrain the search for referents of definite noun phrases. Discussing context, she states in the introduction of the paper,

> A combination of contextual factors influences the interpretation of an utterance. In fact, what is usually meant by "the context of an utterance" is precisely that set of constraints which together direct attention to the concepts of interest in the discourse in which the utterance occurs. Both the preceding discourse context—the utterances that have already occurred—and the situational context—the environment in which an utterance occurs—affect the interpretation of the utterance. For a dialog, the situational context includes the physical environment, the social setting, and the relationship between the participants in the dialog. . . . the task and dialog contexts combine to provide a focus on those concepts relevant

> to the interpretation of utterances in task-oriented dialogs. [Grosz, 1977, pg. 67]

Grosz proposes a representation in the form of partitioning a semantic network, separating out those parts of the network that constitute a *focus space*. The representation separates the explicitly focused concepts from those that are implicitly focused. This separation was motivated by three facts: the fact that many implicitly focused items will never actually be mentioned, the fact that referents are preferentially found among explicitly focused concepts, and the fact that references to implicitly focused concepts produce focus shifts. Thus, Grosz shifted the search for referents from one based on recency to one based on relevance and focus.

In addition, the paper discusses the fact that task-oriented dialogs reflect the structure of the task, which is a crucial contribution. When discussion shifts to a sub-task, the focus changes and the more general task becomes a suspended focus space (Grosz calls them *open* focus spaces). She states that references to a new subtask or to a new parallel or higher task—or to their subtasks—cause a change of focus. At the end of the paper, Grosz points out that focus-shifts can be used to trigger garbage collection of net structure generated to handle the previous focus space. She also points out that her representation can handle multiple points of view (i.e., what the SNePS papers call opaque contexts, in discussing the Telephone Number Problem [Maida and Shapiro, 1982]).

Grosz has also done work [Grosz, 1978] in which she discusses focusing, as an active process engaged in by dialog participants, and its effects on definite descriptions. The examples she used are task-oriented dialogs between an expert and a novice. She mentions the fact that focus movement can reflect task structure. Grosz points out that although speaker and hearer may not share the same knowledge, beliefs, purposes, and so forth, they must both share the same focus. The descriptions of objects used by the novice and the expert may differ considerably, and this may lead to some discussion merely to have each participant know what entities the other is referring to. Grosz points out:

> Two specific problems relating to descriptions are strongly influenced by focusing. From the speaker's perspective, there is the problem of what to include in a description. From the hearer's perspective, there is the problem of what to do when a description doesn't correspond to any known entity, when it doesn't "match" anything. [Grosz, 1978, pg. 98]

In addition, Grosz states that descriptions are influenced by the information speaker and hearer share about the entity being described, the perspectives they have on it, and the use of redundancy. Redundancy appears because the speaker will try to generate a description that will allow the hearer to identify the referent as rapidly as possible. At the other extreme, the description will be no more complex than required because more time will then be lost for the hearer to process the description than will be gained from having the additional detail. I believe she means that overall wall-clock time is to be minimized. When a description doesn't match, the hearer must decide to accept an inexact match (as correct) or else must negotiate with the speaker to resolve the problem. One can see that focusing and definite descriptions mutually influence each other.

In a now classic article, Grosz and Sidner [1986] gave an integrated statement of their influential discourse theory, which accounts for some previously unexplained phenomena of discourse. They postulate three levels of structure. At a surface level, which they call "linguistic structure," there are surface level discourse segments and surface linguistic devices whose direct purpose is to communicate discourse structure. They also postulate an intentional structure of discourse purpose, DP, and discourse segment purposes, DSP: both of these are intended to be recognized, and will not serve their purpose if not recognized. The intentional structure is characterized by sub-purpose dominance and temporal satisfaction-precedence relations. Finally, they postulate an attentional structure of the discourse, consisting of focus spaces and a pushdown-stack structure reflecting the nesting of focus spaces. Pronouns and incomplete definite noun phrases find their referents in the current focus space or the space nearest the top of the stack of focus spaces that contains a suitable referent. Interruptions of various kinds are an example of events that modify the stack of focus spaces. In the design of CircSim-Tutor, we need to incorporate a method of stacking focus spaces, as explained by Grosz and Sidner in this article.

Significantly extending the applicability of the work of Grosz and Sidner, Reichman [1985] studied focus and pronominalization in dialogue rather than expository text. Reichman's work investigates the nature of "context spaces" in discourse and the way in which the context space structure of a discourse determines pronominalization and other surface phenomena. Her work actually *focuses* on multi-participant conversation, although it also applies to expository prose. One of her central findings is that the structure of context spaces determines what can be talked about at any point in a discourse (i.e., what context spaces exist and what sorts of new context spaces are possible). She ties her analysis to Grice's conversational maxims, and in a significant accomplishment showed that implementation of her theory fulfills Grice's maxims.
Reichman also shows that her analysis covers the analyses of Grosz and of Sidner. In fact, her analysis goes beyond the Grosz and Sidner paper [Grosz and Sidner, 1986] in its analysis of what can and cannot be pronominalized. It achieves this by accounting for contextual concerns that Grosz and Sidner do not handle. In essence, Reichman's work points out that the choice of whether or not to pronominalize is determined by, rather than itself determining, high vs. low focus.

It is clear that we need to consider Reichman's analysis in building our text generation. Our discourse generation will have to create some representation of focus/context spaces, and the surface generation must consult this information. Reichman's analysis and implementation are, however, quite complex and detailed. It is not yet clear to me just what of her work we *must* implement, nor have I yet figured out how we might go about it. Her approach and that of Rhetorical Structure Theory (RST), which is discussed below, seem quite different. I believe an RST-based discourse generator will probably create the correct structure, but how to capture it and communicate it to the surface generation is still puzzling. The same comment actually applies somewhat to implementation of the work of Grosz and Sidner [Grosz and Sidner, 1986]. Thus, these aspects of our task are being deferred to future work.

Closely Related Work on Tutoring and Dialogue

Fox [1993] gives a perceptive linguistic analysis of how tutoring dialogue works. Her work includes particularly interesting and relevant analysis of correction in tutoring (chapter 5) and of the effects of the difference in bandwidth between face-to-face and keyboard-to-keyboard tutoring (chapter 8). Clark and Schaefer [1989] discussed how shared context is achieved and accumulates; the research in that paper seems to map well to our transcripts. Wick and Thompson [1992] discussed the problem of computer generation of a "line of explanation" differing significantly from, but having necessary links with, the "line of reasoning" used in problem solving in expert systems. That research appears useful for computer generation of good responses to student initiatives.

Rhetorical Structure Theory (RST)

Rhetorical Structure Theory (RST) was created by William Mann and Sandra Thompson [Mann and Thompson, 1986] [Mann and Thompson, 1987] [Mann, Matthiessen, and Thompson, 1992]. Their 1987 technical report [Mann and Thompson, 1987] gives the most detailed statement of the theory. RST gives a hierarchical tree-structured analysis of a text. The analysis divides the text into meaningful parts called units. Each clause of the text is typically a unit in RST. There are two key points to this theory. First, the relations between units give rise to relational propositions that carry some of the meaning of the text. Second, each unit of the text is assigned one of a set of relations, which concomitantly gives a functional account for why the writer included the unit; and in fact the analysis into a hierarchical structure of relational propositions gives a functional account for the entire text.

RST has four levels. The first level, relations, is the basic building blocks. Relations in RST are approximately the same idea as “rhetorical relations” in McKeown’s work [McKeown, 1985]. Stock RST, as presented in the references just mentioned, has a set of about 20 relations (the set can be customized for different kinds of texts or different applications of RST). At the second level, the relations can be diagrammed as “schemas.” The third level is that schemas can be instantiated, yielding “schema applications.” Fourth, schemas can be composed to yield an analysis of a text as a whole, giving “RST text structures.”

An RST relation typically has a nucleus and a satellite. The nucleus carries the gist of the text, and the satellite is subordinate material. The relation is really between the nucleus and satellite. A relation has four fields: constraints on the nucleus, constraints on the satellite(s), constraints on the combination of the nucleus and satellite(s), and the effect of the relation. The first three fields may be empty, but the effect is required—this is the means by which RST gives a functional account for the structure of the text. As Moore and Pollack [1992] (discussed below) point out, the set of relations in RST is divided into subject-matter relations, which convey information, and presentational relations, which capture intention. As the RST text structures are trees, RST's functional account for the structure of a text must choose between an informational or an intentional account for the function of each piece of the text.

Mann and Thompson make three claims for RST as an accurate model for text: the predominance of nucleus/satellite structural patterns reflecting prominent vs. subordinate material, a functional basis of hierarchy, and a communicative role for text structure itself. It is claimed that RST is successfully independent of English. I have found it straightforward to analyze paragraphs of text in German using RST.

In my opinion, the functional account of text structure that RST gives maps well to our human-tutor transcripts and is more or less what we need as an underlying structural formalism for generating multi-sentence productions in CircSim-Tutor. We really will be working from purpose to text. One point to note is that RST does not establish a fixed set of relations, and I see a need for us to customize Mann and Thompson's set somewhat. I deal with RST more extensively in the chapter of this thesis that deals with my approach to discourse generation and structure.

Alison Cawsey

Alison Cawsey's PhD work at the University of Edinburgh led to an implemented Intelligent Tutoring System (ITS), called EDGE, for simple electronic circuits. She later published a book in which she described this work in great detail [Cawsey, 1992]. The EDGE system carefully considers issues of effective explanation, as well as some important issues of tutorial dialogue. EDGE represents a signal advance in the world of interaction with ITSs.

The student input to EDGE is in the form of menu choices. The menus provide answers to the tutor's questions along with certain questions for the student to ask if desired. The book itself does not satisfactorily convey the limitations on possible questions, which menus obviously must impose. Despite the occasional reminder that student input is actually from menus, the book describes the interaction *as if* the student had typed in his/her questions at the keyboard in English.

The system output is English text displayed on the screen, along with circuit diagrams and other graphics. The English output is generated from templates, customized mainly by filling in NPs, coupled with rather simple pronominalization heuristics. The graphics output, an electronics circuit diagram, often includes a large arrow pointing to the element of the circuit diagram being discussed. In much the same way, a human tutor discussing a circuit diagram with a student face-to-face could point to the diagram. In addition, the graphical display often includes meters showing the input and output values for the components being discussed in the diagram. Cawsey considers issues of how to mix text with graphics in the interaction with a tutoring system.

Although EDGE is a tutoring system, it really focuses on interactive explanations rather than the multiple types of text delivered by CircSim-Tutor. What is perhaps *the* central issue for EDGE is how interactions with the user can be managed within a complex explanation, and how they should influence the way that explanation develops. Cawsey's description of how EDGE works internally, begins with lengthy treatments of the major tasks the EDGE program must perform. Tasks include choosing and organizing content, organizing the dialogue with the user, updating the user model, and the circuit domain specifics. Particularly noteworthy in Cawsey's description of her EDGE system is that she gives an extensive evaluation of it. This evaluation also includes a very interesting discussion of ideas about how to eliminate the main shortcomings of EDGE. The evaluation apparently reflects extensive user experience, as the list of shortcomings is long and includes rather subtle issues. Cawsey states,

> The system succeeds, in a fairly simple and restricted way, in generating coherent, individualized, responsive, and robust explanations encompassing a range of dialogue phenomena,. . . . Unlike most dialogue systems the system views the planning of an interactive explanation as an extension of planning a non-interactive explanation. . . . Dialogue and monologue are described in the same framework. [Cawsey, 1992, pg. 126]

The planning in the EDGE system strikes me as largely organized around discourse issues and user-model issues (in contrast to the proposed CircSim-Tutor organization around curriculum/content planning and user-modeling issues).

EDGE is the expert-dominant party in its tutoring sessions. When the student interrupts with a question, EDGE will try to return to its agenda, but will also try to maintain the focus introduced by the student if convenient. In addition, during an interruption the user model may be updated. Thus, an interruption may change both the content and discourse structure of whatever follows the return to EDGE's agenda after an interruption.

Cawsey points out [1992, pg. 11]:

> Text planning is concerned with how a coherent, appropriate, multi-sentence text may be generated given some communicative goal. The communicative goal may be of various kinds, such as:
> - to **describe** an object.
> - to **convince** the hearer to take some action.
> - to make the hearer **believe** or **understand** something.

Thus Cawsey's view of text is basically that it consists of speech acts [Austin, 1975; Grice, 1975; Searle, 1969, 1975].

In discussing user models, Cawsey states that

> A common approach (used in the EDGE system) is to start with some *stereotype* based model, where the user's knowledge can be guessed from their general class or level of expertise. This model can then be updated and refined based on particular interactions with the user. [Cawsey, 1992, pg. 12]

"Stereotype based" means that you classify the user and have default assumptions about what a user in each class knows. In fact, the EDGE system makes heavy use of user stereotypes to guide interaction with the user, for example in deciding what to omit and what to include in an explanation, or in deciding what topics to explore.

Cawsey's analysis of human explanations has three types of categories: content of the utterance, role of the utterance in the whole text, and type of explanation (e.g., using *analogies* or *examples*). In more detail, Cawsey's categories used in analyzing explanations (her Fig. 2.3, pg. 21) are as follows.

<u>Type of Role/Rhetorical Relation</u>:

* Background	* Summary/conclusion	* Elaboration
* Compare-contrast	* Sequence	

<u>Type of Content</u>:

* How-it-works	* Structure	* Process
* Behavior	* Identification	* Constituency
* Function	* Causal-event	* Component(s)

<u>Type of Explanation</u>:

* Examples	* Classification	* Analogy

Note that her "Type of Role/Rhetorical Relation" category has the same ideas as the more comprehensive analysis in RST. As Mann and Thompson [1987] themselves point out, McKeown's schemas are a finer-grained analysis of the RST *elaboration* relation. Referring to RST, Cawsey states,

> . . . in our experience such relations on their own provide an insufficiently fine-grained description of text structure not easily adapted to generating particular types of text in a particular domain. . . . We therefore base our analysis and text planning primarily on a simple content grammar, with rhetorical relations playing a less prominent role. [Cawsey, 1992, pg. 26]

Drawing on Conversation Analysis, Cawsey sees dialogue as having multi-turn structures consisting of "exchanges." Boundary exchanges begin both major topics and subtopics, so Cawsey introduced the idea of a transaction on a major topic consisting of more fine-grained transactions on subtopics. She views an informing transaction as consisting of individual informing and teaching exchanges, rather than a single informing exchange (transaction?) with embedded teaching exchanges. She notes that lack of interruption by the student can be viewed as a kind of implicit acknowledgement [Cawsey, 1992, pg. 31]. Previous researchers had found that an informing transaction begins and ends with a boundary exchange (an opening exchange or closing exchange), but Cawsey found that the final informing transaction was not closed, thus leaving the initiative with the student. Our human tutors certainly do this. Cawsey found that, often, the end was a pause or asking the student whether it was OK to move on to the next topic—thus a monologue by the teacher actually has embedded acknowledgements by the student [Cawsey, 1992, pg. 32]. Closing exchanges in EDGE have a basic pattern in which the system suggests closing an explanation of some topic, and the user acknowledges. Cawsey found that feedback from the tutor may become an entire remedial subdialogue when the student makes an error [Cawsey, 1992, pg. 36]. Similarly, answers to students' questions may be elaborate subdialogues. In general, Cawsey found that earlier work on Conversation Analysis required extensions to handle *interruptions* in the dialogue [Cawsey, 1992, pp. 36–37].

Finally, Cawsey found two kinds of exchange apparently not described earlier. One is that either teacher or student can request the other participant to alter the style of presentation. The explanation being given by the teacher, or the answer being given by the student, may need to be supplemented ("Please write down formulas so we can see what is going on." Or, "I need to know the point of this particular bit rather than an explanation at a lower level.") The other additional kind of exchange is that the teacher may simply ask the student whether some fact is known/remembered, rather than asking a question testing the student's knowledge. While we are all familiar with teachers and tutors asking students whether they remember some fact, our tutors assert that asking students whether they remember something is not an effective way to find out whether they, in fact, actually do remember it. One must get the student to state the material, or better yet get the student to *use* the material successfully, before one can be sure the student knows and remembers it.

Cawsey points out that the tutor may set the agenda in a tutorial dialogue, the student's questions may set the agenda, or there may be a joint/equal relationship in setting the agenda. Also, the tutor may or may not verify/test the student's understanding. In CircSim-Tutor, the tutor always sets the agenda.

Cawsey also takes up the relation between content organization and dialogue structure. She points out that the relation between dialogue and content structure is analogous to the relation that Grosz [Grosz, 1977] found between dialogue and task structure. Cawsey credits Woolf and Murray (in a paper in the 10th IJCAI, 1987) with suggesting that schemata can capture common remediation strategies. Cawsey states

> In their remediation schema there should be some "grain of truth acknowledgement" followed by one of three types of remediation depending on the type of error. These types of remediation involve teaching the students the *consequence* of their incorrect answer, or giving *examples* or *guidance* leading the student toward the correct answer. [Cawsey, 1992, pg. 46]

In EDGE, remediation strategies "may depend on the type of question asked, on what the user knows, and on the particular pedagogical or dialogue strategy being pursued." [Cawsey, 1992, pg. 47]

Analyzing one of the dialogues in her corpus [Cawsey, 1992, pg. 46, Figure 2.21], Cawsey states:

> Two strategies are apparent. If a question is answered with a contributing step (rather than the final answer) then the expert acknowledges that step and prompts for further steps. If the answer to the original question is incorrect or confused, then the expert may use the strategy of prompting for substeps to lead the novice to understand the answer. Another strategy seen in [Cawsey, 1992, pg. 43, Figure 2.19] is to supply supplementary information (such as substeps or an analogy) and then ask the question again. [Cawsey, 1992, pg. 47]

The EDGE system makes a distinction between planning explanation content and managing the dialogue with the user. The user model affects both. In discussing the planning of explanation content, Cawsey lists eight principles governing the structure and content of texts. Cawsey states that DeBeaugrande and Dressler proposed the first seven of these as standards of *textuality*. [Cawsey, 1992, pp. 54-59]

- Cohesion: surface level ties (between concepts or ideas)

  e.g., pronominalization, conjunctions, parallelism, ellipsis

  Pronominalization is essential if the user is to track focus.

- Coherence: conceptual relations

    the reader must grasp the conceptual relations underlying the text

- Intentionality: the speaker's goal

    e.g., Grosz&Sidner's dominance and satisfaction-precedence relations

- Acceptability: the hearer's goals and attitudes

    the readers must find the text to be relevant to their goals

- Informativity: the hearer's knowledge

    neither too much nor too little, and with needed background info

- Situationality: the discourse context

    both the preceding part of the discourse and the overall wider situation in which the discourse occurs (how to be relevant to the context)

- Conventionality

    Often there are fixed conventions for the type of discourse.

- Extralinguistic devices

    e.g., graphics, sound effects, and pointing actions.

Cawsey discusses the need to *plan* text content rather than using schemas.

> As long as we are only concerned with a small range of text types there isn't too much overhead. However, if we want to generate texts which depend on further and more detailed aspects of the discourse situation then the approach appears to suffer from the same problems as the canned text approach. . . . In order to generate texts to answer seven types of question, addressed to three classes of user, we might in principle want to create up to 21 schemata, each obtained from the detailed analysis of many human textual descriptions of the appropriate type. [Cawsey, 1992, pg. 61]

Cawsey's comments on schemas are somewhat the same as my doubts about capturing the internal scripts that our human tutors say they follow.

Cawsey makes some comments on planning informative exchanges.

> Generating an understandable explanation must involve, at the least, choosing between alternative ways of explaining something and choosing when additional background material should be provided. . . . To decide on background material, prerequisite relationships may be defined among topics, specifying explicitly what topics must already be understood for others to be successfully explained and understood. . . . the prerequisite topics don't just depend on the topic itself but on the way it is to be explained.
>
> . . . An explanation strategy may have an associated constraint, indicating the class of user the strategy is likely to be effective (and efficient) for. At the simplest, the constraint may simply indicate that this strategy is likely to be good for novices. . . Slightly more complex, a constraint might indicate that a particular strategy (e.g., *analogy*) is likely to be good for people who know about the analogous object. [Cawsey, 1992, pp. 65–66]

Explanations are often improved by omitting prerequisite or background information that the hearer already knows. However, sometimes information plays a different role in an explanation. When explaining a process that has a series of cause and effect steps, the explanation can be incoherent if steps are omitted. Thus, one cannot decide to omit information solely because the hearer already knows it. One must consider which of these two roles, background vs. essential step, the information plays in the explanation. Although the choices may be a bit hard to make, Cawsey's content planning rules distinguish the two.

Preconditions in her rules identify prerequisite or background information which needs to be included (if not already known) in order for the “main” explanation to make sense. For example, knowledge of the behavior of a device's components helps the user to make sense of the explanation of how that device works, but need not be explained if the user already understands it.

On the other hand, subgoals represent the crucial subtopics in the main explanation. These subtopics will be included even if the system believes they are already understood. It is often the sequence in which information is presented that conveys new conceptual information (i.e., important relationships between known concepts), so it is important not to omit such topics or concepts even if they are already understood [Cawsey, 1992, pg. 68]. Cawsey states that EDGE has approximately 25 content planning rules.

The EDGE system uses discourse markers and meta-comments primarily to indicate a new topic in the discourse. A topic is introduced by a topic opening discourse marker (e.g., “Now” or “Alright”) and meta-comment(s) on what is about to be explained. An opening utterance resulting from this is, “Right, I'm going to explain how the light detector works.” [Cawsey, 1992, pg. 82] As mentioned above, in closing exchanges the system suggests ending an explanation of a topic and the user indicates that closing the topic is acceptable.

The general structure of a teacher-response move in EDGE consists of a Yes or No, followed by some expression of the teacher's opinion of the student’s answer, and then a factual comment. Here are some examples.

> “Yes, very good, it's high.”
>
> “No, not quite, it's actually quite low.” [Cawsey, 1992, pg. 108]

In dealing with lengthy interruptions, a tutor needs a way to resume the interrupted topic. One observed technique is to say, “Anyway,” possibly followed by a reminder of the previous topic. For example, “Anyway, we were in the middle of explaining how the light detector works.” A particularly common tactic in verbal explanations is to repeat part of what was said before the interruption, for which Cawsey gives the following example. [Cawsey, 1992, pg. 117]

Expert: Each of these acts rather like a switch, so if this is high . . .

Novice: You say shorted back to base, but there's a capacitor?

Expert: Right, there is indeed, coming to that in a sec.

Each of these is acting like a switch, so if this voltage here is. . . .

Major subtopics in a complex explanation should also have markers and meta-comments. To handle this, EDGE allows *subtransaction* subgoals in the content planning rules. These may include an argument indicating the functional role of the subgoal. EDGE introduced this extension because the uniform top-down planning makes it hard to predict which subtopics will turn out to be "major." Cawsey notes that writing rules really presents the same problem, predicting in advance, that the top-down planner faces. Some objects are more complex to explain and some users know more than others do. These factors may greatly affect the complexity of a subtopic. In addition, "marking functional roles in this way is a less elegant solution than approaches such as Moore's and Hovy's based on Rhetorical Structure Theory." [Cawsey, 1992, pg. 84]

Summarizing, Cawsey states:

> In the long term we need to consider how domain-independent dialogue systems can be developed which take into account both the plans and intentions of the participants and conventional practices. These conventional conversational practices (such a methods of turn taking and conversational closings) are designed to enable local, context-dependent, collaborative decisions. They therefore cannot be ignored. It is unlikely that they will all be reduced to complex intention-base decisions, and even if they could, for current practical

> systems it is useful to explicitly represent conventional conversational structure. The EDGE system uses a primarily "convention-based" dialogue grammar (represented as simple planning rules) to manage the interaction with the user. [Cawsey, 1992, pg. 100]

Cawsey's system has a clear separation between content planning rules and those that manage the dialogue. I believe this is an important accomplishment of her system.

Studies of Questions and Answers

Graesser, Lang, and Horgan [1988] proposed an analytic scheme for questions, covering a corpus of approximately 1,000 questions asked by adults in different discourse contexts. They proposed 12 semantic categories for questions.

- Verification: Is X true or false?
- Disjunctive: Is X or Y the case?
- Concept completion: Who? What? When? Where?
- Feature specification: What is the value of a variable?
- Quantification: How much? How many?
- Causal antecedent: What caused some event to occur?
- Causal consequence: What happened as a consequence of X occurring?
- Goal orientation: Why did an agent do some action?
- Enablement: What is needed for an agent to do some action?
- Instrumental/procedural: How did the agent perform an action?
- Expectational: Why isn't X occurring?
- Judgemental: What should an agent do?

Graesser, Lang, and Horgan [1988] also proposed 6 pragmatic categories, intended to be orthogonal to the semantic categories. These categories are: information acquisition, assertions, establishing a context for subsequent discourse, indirect requests for non-verbal behavior, conversation monitoring, and humor.

Graesser and Franklin [1990] proposed a cognitive model of question answering, called QUEST. This model specifies the information sources that furnish answers, the formal representation patterns that could account for the structure of that information, and convergence mechanisms that find the relevant information to include in an answer. The information representation is structured as a graph, with nodes and arcs/edges. Graesser and Hemphill [1991] studied the convergence mechanisms of QUEST by doing experiments in which the subjects read brief texts in various domains and answered questions about them. Each text presented a causal chain with five steps. The questions dealt with specific pieces of information present in the text: why, when, how, enablement, consequences [Graesser and Hemphill, 1991, pg. 187]. The three domains were physical science (tornados, earthquakes, light, rain, sonic boom, riptide, supernova, stalagmite formation), biology (seeing, the heart, photosynthesis, knee jerk, mitosis, hair growth, hearing, neurons), and technological systems (computers, television, paper production, nuclear energy, elevators, vacuum cleaners, water purification, wine production) [Graesser and Hemphill, 1991, pg. 194]. They reached the following conclusions. The questions that are easiest for people to answer have answers that are close to the material asked about, by which they mean that one has to follow a very short path in the information graph to get from the node in the information base that represents the question to the node that represents the answer. In fact, they also showed that their subjects rated answers as being better if the answers were closely related to the questions in this sense, and that raters arrived at goodness-of-answer judgments more rapidly if the answer was closely related to the question. If we cast these results in terms of a causal chain, the ideal was for the answer to be found only one causal step (either antecedent or consequent) from the event asked about [Graesser and Hemphill, 1991, pg. 198].

Even more interesting for us was their study of "why" questions in the three domains. It turned out that when answering questions about texts in biology and technological systems, there was a clear preference for answers that stated causal consequences. In other words, the underlying human reasoning process of choice was to see the causal chains in the biological and technological domains in teleological terms (i.e., in terms of the *purposes* they serve rather than in terms of their *causes*). In contrast, in the physical science domain, the preferred answer to "why" questions was a statement of causal antecedent(s), which means the preferred human reasoning technique in this domain is purely in term of cause and effect.

What makes this finding particularly interesting for CircSim-Tutor is that the baroreceptor reflex, which CircSim-Tutor is intended to teach, falls in the biology domain, but the necessary style of reasoning is purely in terms of physical cause and effect. The few students in the transcripts who did reason teleologically went badly astray in the process and had difficulty understanding the actual causal chain of events because the actual causal chain tended to conflict with their underlying teleological model. Of course, one can actually recast the physical cause-and-effect chain of the baroreceptor reflex in teleological terms and reason correctly, but the students appear far more likely to have an incorrect teleological model. Earlier, I discussed the role of schemas (in the cognitive psychology sense of the term) in making it difficult for students to accept what they were being taught by the tutors. This is a somewhat related problem, since one can easily view the problems experienced by students who reason teleologically as being due to having a teleological schema for the baroreceptor reflex. When our tutors discover the student is reasoning teleologically about the individual steps in the causal chain (and doing so incorrectly) they respond by telling the student explicitly that there is no purpose for the individual steps in the chain of cause and effect. The body senses the mean arterial pressure (MAP), MAP has a fixed setpoint value, and when MAP rises or falls (departing from the setpoint value) the baroreceptor reflex acts to restore MAP toward the setpoint. Thus, the students are probably used to reasoning teleologically about biological systems, and they need to be taught, instead, to reason purely in terms of physical cause and effect.

Work by Johanna Moore

Johanna Moore's [1995] work was done for the Explainable Expert System (EES) project. Other researchers who have worked on that project are Cecile Paris and William Swartout. Moore points out that important tasks for an Expert System (ES) include defining terms, justifying results, comparing alternative methods of solving problems, and handling follow-up questions. Studying the generation of good responses to follow-up questions is an important aspect of Moore's work. Such questions arise when a user did not understand an explanation or an answer to a question, when the user wants more information (e.g., "Why?" or "What's that?"), and when the user wants to ask about some part of an explanation. The expert system must understand the intention behind the question, remember the intention behind its response, and must be sensitive to context (current task and preceding dialogue).

A crucial contribution of Moore's work is to identify the kinds of knowledge needed to handle follow-up questions correctly. Moore's planner explicitly represents four kinds of such knowledge:

- *wha*t is being conveyed in an explanation (content),
- *why* each part is being said (intention),

- *how* it is being conveyed (rhetorical strategy), and
- *what assumptions about the user* are being made (user model).

When the user fails to understand, or asks a follow-up question, any one of these may be needed to respond appropriately. Moore states that other researchers have concluded that advisory dialogues are a process, in which the participants *negotiate* the problems to solve and the solutions that the advice seeker can understand and accept.

Interaction with an expert system becomes complicated because users frequently do not understand the expert's response, users frequently ask follow-up questions, users often do not know *what* they do not understand, and expert do not have a complete or accurate model of the user.

Better explanations are enabled by improvements in expert system and knowledge representation (KR) technology. With frame-based KR systems, the meaning of the knowledge stored in the KR system springs from the system's algorithms and the data-structures that the algorithms work on. In contrast, KR systems in which the meanings of represented concepts are unambiguously determined by *explicit notational devices* whose meanings are well-defined and understood (*criterial* semantics) enable better definitions of terms and types of knowledge. However, translate-the-code approaches are more difficult with KR systems based on criterial semantics.

In discussing text generation by graph traversal, Moore points out that graph traversal is the right approach when discourse structure mirrors task structure. The work of Grosz [Grosz, 1977] [Grosz, 1978] described earlier is a good example. Graph traversal may involve causal, spatial, or temporal graphs. Moore points out that focusing is needed in graph traversal. Later in this dissertation, I will explain how I generate summaries of the direct response phase of clinical problems in CircSim-Tutor. Those summaries are partly generated by graph traversal.

In discussing text generation from schemas, Moore points out that schemas give domain-independent explanations and decouple the explanation strategies from knowledge representation. In fact one can generate a range of explanations from a single knowledge base. However, intentions are not represented in schemas (except sometimes as the top-level goal) and opportunistic effects can be difficult to achieve (e.g., new terms are introduced everywhere in text, and it would be helpful to define them opportunistically when they first appear).

Much of Moore's work concerns the effective representation of the intentions behind an explanation or answer, and the use of her techniques of representing intentions when generating text in a (basically) RST-based theory. I am not generating text using an RST-based approach, and I have not yet been able to properly exploit her work in this area. Since I am not yet using it, I am going to omit a detailed presentation of this part of her work. However, by representing intentions in detail and being able to consult this representation when follow-up questions ensue, Moore is able to try different approaches to achieving the original goal. I find this to be an exceedingly impressive accomplishment indeed. It is built on her system's ability to reason about its previous utterances. Moore's work has obvious applicability to dealing with repair and in responding to student initiatives. For us to make use of what she has achieved, I think we will need a much more rule-based approach than the rather procedural approach I am currently using. Thus, the most important gains that we could take from Moore's work must, alas, be deferred to future work.

Previous Work in the CircSim-Tutor Project

Previous research on the earlier transcripts in our corpus of keyboard-to-keyboard transcripts has been published by Seu, Chang, Li, Evens, Michael, and Rovick [1991], Hume, Michael, Rovick, and Evens [1993, 1995], and Evens, Spitkovsky, Boyle, Michael, and Rovick [1993]. Seu, Chang, Li, Evens, Michael, and Rovick [1991] investigated the differences between face-to-face and keyboard-to-keyboard tutoring, showing among other things that there are syntax changes and that keyboard-to-keyboard tutoring uses a larger English vocabulary than is used face-to-face. The differences appear to fall out from the difference in bandwidth of face-to-face vs. keyboard-to-keyboard communication.

CHAPTER III

STUDENT INITIATIVES

Introduction

By student initiatives we mean productions by which the student is apparently trying to modify the course of the tutorial dialogue and could reasonably expect to do so. Asking a question is one kind of student initiative. Here are some typical examples (partly invented) of student initiatives. The tutor's turns are marked "tu" and the student's turns are marked "st." The actual initiatives are underlined.

tu: What happens next?
st: I think that with Heart Rate increased, the force of contraction goes up. <u>But I don't know how that would be true.</u>
. . .
tu: You have predicted HR i, CO i and SV d. How do you explain this?
st: <u>What do you want me to explain?</u>
tu: Can you write an equation relating HR, CO, and SV?
st: Yes. CO = HR x SV. <u>In this patient, SV is determined by the CO.</u>
tu: Correct. So what happens next?
st: Right Atrial Pressure falls<u>???</u>
tu: You're not sure?
st: <u>Why wouldn't it rise with increasing venous return?</u>

In order to make CircSim-Tutor handle the discourse phenomena found in the sessions with human tutors, I set out to analyze the transcripts of keyboard-to-keyboard tutoring sessions in which the students were first-year medical students (the target users of CircSim-Tutor) and the tutor was Joel Michael or Allen Rovick (each did about half the sessions). During these sessions, the student was in one room and the tutor in another room, communicating only by typing on the keyboard and reading from the screen of a computer terminal. Each student always knew that the interaction was with a human tutor and also knew the identity of the tutor. In addition, each student also had one of the tutors as his or her professor in the related physiology course being taken at the same time as part of his or her first-year medical school coursework at Rush. Thus, the students may have felt the academic or social pressure this would imply. All students were volunteers, recruited from these classes by flyers that promised participants that they would learn some useful physiology, and each was paid a nominal amount for his or her participation.

In most tutoring studies in the literature, the students are seeking remedial help with learning material and skills with which they are having trouble. Our students did not yet know whether they needed help with the material. These sessions occurred at a point in the students' physiology coursework where they had had the opportunity to learn the necessary background facts. They had heard the class lectures or equivalent on the material, had probably read the textbook material on the topics, had had a laboratory on normal cardiovascular physiology, and had had a small-group discussion session. But, the students had not yet had any exams or clinical problem sessions that would show them that they personally need help on this particular material and skills. Thus, we believe that all the students in these sessions were in the process of initially learning how to apply their knowledge of the material in solving clinical problems, and that none were participating in order to get remedial help. As has been stated, the focus of these sessions was learning to solve clinical problems. CircSim-Tutor will be used at the same point in the physiology course in the students' first-year program at Rush Medical College and for the same purposes as the students were participating in the tutoring sessions we studied.

We have obtained demographic information on 18 of the 19 students in the second set of sessions analyzed here: 8 are male, 10 are female. Their ages range from 21 to 37, averaging 26.25 years old. Almost all speak English with native fluency. They are typical of students in American medical schools. We believe the students in the first set of transcripts are demographically similar, even though we did not collect the specific demographic information.

The basic agenda of the tutoring sessions is to present clinical scenarios that would directly or indirectly perturb the arterial blood pressure. In order to teach inference and problem solving skills that are based on thinking causally, the tutor then asks the student to predict (in detail) what will happen and to explain certain interesting predictions. Tutoring typically centers on this discussion of predictions and reasons for them. It may also take off from student initiatives.

The clinical problems represented in these sessions were as follows. A person has an artificial pacemaker which is the sole determinant of the heart rate, and the pacemaker suddenly fails, increasing the heart rate substantially. A patient has a liter of whole blood rapidly infused intravenously. A patient is given an alpha-adrenergic agonist drug with a dose sufficient to produce the maximal alpha-agonist effect. A patient is given the opposite sort of drug: a maximal dose of an alpha-adrenergic antagonist. A person is put into a human centrifuge with the g-force directed toward the feet and away from the trunk and head.

I first analyzed a set of transcripts of 28 hour-long keyboard-to-keyboard tutoring sessions. In these sessions, Joel Michael and Allen Rovick tutored a total of 20 students (8 students appeared twice). I later analyzed a second set of 19 sessions that lasted closer to two hours each and in which they tutored a total of 19 additional students. The volume of this second set transcripts is approximately equal to the volume of the first 28. As the transcript of one of the 19 new sessions is incomplete, we have omitted it from our analysis.

In this chapter, I give a classification of the student initiatives found and the tutor's responses to them, and discuss examples, with particular attention to the thinking that appears to underlie the tutors' responses.

Joel Michael and Allen Rovick, the two Physiology Professors at Rush who are the tutors in these sessions, have taught the related physiology courses to first-year medical students at Rush for many years, and have customarily tutored some students taking these classes face-to-face. Thus, our tutors are highly experienced at teaching the material, both in classroom lecture and in personal tutoring sessions. Subjectively, they also seem extremely expert at tutoring this material.

Our tutors are also experienced and expert about the context in which these sessions occurred. They all occur at the same point in the students' physiology coursework in their first year of medical school. Because our tutors are experienced Physiology professors in these students' program, they are experts about this context. Specifically, our tutors know what it is most important that the students learn from the sessions (i.e., what they need to learn at this point in their study of physiology), and what difficulties the students typically do and do not have at this point in their study. So, our tutors come to these sessions with pre-existing overall tutoring goals and considerable insight into the students.

Transcripts including timing information were collected automatically using a program called CDS [Li, Evens, Michael, and Rovick, 1992]. The basic scenario in these tutoring sessions was as follows. Each of the 28 hour-long sessions was organized as a clinical problem, where a mechanical heart pacemaker suddenly failed, increasing the heart rate (beats/minute) substantially. Each of the later set of 19 longer sessions covered one or two different clinical problems that perturbed the blood pressure. This is stated in more detail below. In solving the problems, the student was asked to predict the direction of change, if any, of seven basic cardiovascular parameters, first for the immediate physical effects of the perturbation, then for the reflex compensation by the autonomic nervous system to return the blood pressure toward the original value, and finally for the steady-state result after this compensation is complete. In addition to making the correct predictions, the tutors want the students to be able to explain why and how each of the changes occurs, and to do so using the "correct" language.

The number of initiatives per hour-long session in the first set of transcripts ranged from 0 to 11, with a standard deviation of 3.1 and mean of 3.9 per session. Let me give some more detailed numbers about the content and initiatives in the second set of tutoring sessions. A total of 18 sessions from the second set are analyzed here. Seven of these sessions covered one clinical problem in each session, and the other eleven sessions each covered two clinical problems. Each session lasted until the material had been covered: the single-problem sessions lasted an average of 79 minutes each, and the two-problem sessions lasted an average of 113 minutes. The 18 sessions lasted a total of 29 hours 55 minutes. The sessions in which Joel Michael was the tutor lasted a total of 12 hours 53 minutes. Those in which Allen Rovick was the tutor lasted a total of 17 hours 2 minutes. There are a total of 150 student initiatives in these sessions: 60 in the sessions with Joel Michael, 90 with Allen Rovick. The maximum was 17 initiatives in a session and the minimum was 1. On average, there were 8.33 initiatives per session, one every 19.8 minutes with standard deviation 15.7 minutes. The session with the most frequent initiatives had one every 6.9 minutes. At the other extreme, the session with only 1 initiative lasted 64 minutes.

Identification and Classification of Initiatives

As stated earlier, I analyzed the transcripts of the keyboard-to-keyboard tutoring sessions done by our human tutors. We identified and categorized each instance where the student took the initiative in these sessions and described how the tutor responded to the initiative.

As mentioned in Chapter II, Graesser, Lang, and Horgan [1988] studied a corpus of approximately 1,000 questions asked by adults in different discourse contexts and proposed 12 semantic categories for the questions that they found.

- Verification: Is X true or false?
- Disjunctive: Is X or Y the case?
- Concept completion: Who? What? When? Where?
- Feature specification: What is the value of a variable?
- Quantification: How much? How many?
- Causal antecedent: What caused some event to occur?
- Causal consequence: What happened as a consequence of X occurring?
- Goal orientation: Why did an agent do some action?
- Enablement: What is needed for an agent to do some action?
- Instrumental/procedural: How did the agent perform an action?
- Expectational: Why isn't X occurring?
- Judgemental: What should an agent do?

In analyzing the transcripts with Joel Michael and Allen Rovick, we seemed to need an added category: questions about ontology or taxonomy.

Graesser, Lang, and Horgan [1988] also proposed six pragmatic categories, intended to be orthogonal to the semantic categories. These categories are: information acquisition, assertions, establishing a context for subsequent discourse, indirect requests for non-verbal behavior, conversation monitoring, and humor. These may indeed cover the questions we found in our transcripts.

The following is my classification of the student initiatives. Although primarily semantic or pragmatic (generally, discourse-structure based), some of the categories pick out surface clues that seem to flag a production as an initiative. I defined eight classes of student initiatives, and the classes are intended to make distinctions that are important for the tutor's responses. Most of the classes are subdivided into subclasses, but the subclasses are really intended to help people to understand and use the classification system; the sub-classification have no real theoretical significance. This classification was created from study of the keyboard-to-keyboard tutoring sessions.

- Class 1: The student asks a question.

  In more detail: The student (at least in effect) asks a question about the subject matter (physiology) or about himself/herself, and the initiative is *not* primarily hedging.

- Class 2: The student is having trouble "seeing" something or another (the student is *not* mainly requesting repair)
- Class 3: The student requests repair (the student did not understand the tutor)
- Class 4: Do repair (the tutor did not understand the student)
- Class 5: Hedging by the student
- Class 6: Explicit backward reference to some earlier topic, event, time, etc.

  The reference could be to something talked about earlier in the current session ("Could we talk about *X* some more?) or it could be to previous sessions, conversations, class sessions, etc.

- Class 7: Initiatives specific to the keyboard-to-keyboard environment used in these sessions (asking about something that is now off the screen)
- Class 8: Administrivia

Now, in fact these classes do not lend themselves to unambiguous classification. Interrater agreement is only around 70 to 80 percent. When I discussed their judgments with various raters, the raters clearly disagree about what is actually going on in the student's mind in many of the specific cases of disagreement about what class to pick. This seems to confound the results of my study of interrater agreement. However, the raters also found it difficult to distinguish class 1 from classes 2, 3, and 5 in the case of certain specific initiatives in their actual context. Based on discussion of specific initiatives with raters, I believe the confusion between classes 1 and 2 might lessen with more extensive training of the raters. Much more importantly, the raters all seemed to believe the distinctions between classes are valid as a description of what is going on in our transcripts, even if their classifications of the initiatives did not agree as closely as I expected. It is important to remember that the purpose of the classification is to describe (1) what sorts of initiatives occur and (2) how the students actually do take the initiative. Thus, I continue to believe the set of classes is valid and useful despite the disagreements among raters.

Viewed on a surface level, I found our tutors responded to the initiatives in the following ways.

- Explain or state some material in focus.
- Defer handling the initiative: perhaps modifying the tutor's model of the student. Sometimes, these responses were brushing off the initiatives, sometimes merely deferring the response.
- Do repair, stating some material, where the student did not understand the tutor.
- Request repair: the tutor doesn't understand what the student means.
- Ask the student if he or she is stuck, or still stuck.
- Acknowledge the student's understanding is correct, or state it is not correct.

- Replan part or all of the remaining session:

    perhaps cover material in pieces,

    perhaps make a big backward reference.

- Give a hint, or perhaps remind student of material already covered in the session.
- Ask the student a question. (Socratic tutoring)
- State, "You are confusing *X* with *Y*." (Declare a diagnosis)
- Invite the student to review his/her thinking with the tutor.

In many respects, this classification of how the tutors responded is not particularly interesting, since the real question is what the tutor hoped to accomplish rather than the surface appearance of the immediate response. Nevertheless, it is relatively easy to classify the tutors' responses using these categories, and they do seem to cover the responses seen.

One of the first things we noticed in the transcripts of the sessions is that the students may use punctuation, if at all, in a personal way, often with minimal relationship to the generally accepted conventions of English punctuation. Thus, punctuation may provide little help in recognizing the mood or clausal structure of sentences. We believe most students are simply unable to use punctuation according to the conventional rules while thinking their way through the clinical problems. In contrast, the students generally capitalize conventionally.

Repeated punctuation (e.g., "???" or "!!") always appeared significant. Generally, surface clues are what seem to trigger recognition of a student initiative and of its meaning. The Hedging class (class 5) has some particularly clear examples of this. It appears the students consistently flag all initiatives in some fashion so the tutor does not have to notice a departure from the current discourse focus or make similar inferences to recognize initiatives.

Interesting Examples of Initiatives

All examples are given with the original spelling errors, punctuation, capitalization, typographical errors, and so forth. The following example came at the end of discussing the direct physical effects before the reflex kicks in. The abbreviations used by the tutor and student in this example are: cc=cardiac contractility, tpr=total peripheral resistance, co=cardiac output, ans=autonomic nervous system, ca=calcium [ions], and i=increase. Note that the student flags the material he wants the tutor to respond to by saying, "I'm not sure if. . . ." As mentioned above, students in our sessions consistently flagged initiatives by doing something of this sort.

> tu - One last question here. . . Why did you predict that cc and tpr would be unchanged?
> st - Tpr is largely a function of arteriol constriction which takes a while to adjust to co i .
> st - Cc changes in response to ans stimulation or ca build up during tachecardia.
> st - Im not sure if 120bpm is fast enough to cause that.
> tu - Probably not.

The following is another example, starting in the middle of a tutor's production. The only abbreviation is RAP=right atrial pressure.

> tu - [. . .] what about the rate at which blood is being removed vfrom the central blood compartmanent?
> st - That rate would increase, perhaps increaseing RAP???

In our sessions, the tutors appear to have a well defined picture of what they want the student to demonstrate and what the student should be tutored on if the student does not already know. Interestingly, the mere mention by the student of certain terms not introduced into the session by the tutor is enough to trigger tutoring on the parallels between those parameters and the ones the tutor is using in this session. The parallel in the following example is one of similar values: CVP and RAP are really separate measurements, but their values should be essentially identical. The abbreviations here are CO = Cardiac Output, RAP = Right Atrial Pressure, and D = decrease.

st - So, when CO I, the central venous pressure will D?
tu - Absolutely correct.
tu - What variable is essentially the same as central venous pressure?
st - RAP.
tu - Right.

Some initiatives are quite brief, their interpretation clear, and the response is fairly obvious.

tu - OK?
st - No

Others are complex. In the following example, SV = stroke volume. The student in this example had previously produced a 209 word response to a question, which the tutor eventually interrupted to tell the student, "you need to be more concise in your answers."

tu - Understand?
st - Not fully.
st - Isn't the amount of filling equivalent to the preload?
st - And doesn't and increased preload invoke Starling's effect?
st - And, most importantly, what is the difference between a length/tension effect (as occurs in Starling's) and the "change in ventricular performance (SV, force, . . .)" which you say is not related to Starling.

As has been pointed out by research on discourse or dialogue structure, there is always some current focus, often a nested stack of subjects in focus. The preceding example establishes material in local focus, and the tutor responded by tutoring the pieces separately, then returning to the previous course of the session. This question of whether something is in or out of the current focus, seems important in recognizing the intent of student initiatives and in deciding how to handle them. Our tutors often responded to straight questions that were off the current topic in the briefest possible fashion and then simply returned to the previous topic with no surface flagging that the topic was changing back, as if the focus had never changed. For example, an initiative as long and complex as the preceding example got the response, "Yes." On the other hand, questions about the material currently in focus generally got more elaborate treatment. For example, the following initiative took four st/tu pairs of productions to be discussed. It became a significant topic in its own right, even though this question is not part of the "standard" material to be covered in these sessions. (RAP = Right Atrial Pressure, CO = Cardiac Output, and I = Increase)

> st - Does RAP increase initially with increasing CO and then taper off as CO continues to I?

The most frequent sort of initiative was the subclass of class 1 consisting of straight questions about material currently in focus (about 18%), and initiatives in class 2, trouble seeing/conceptualizing/grasping something (about 7%). Four other common sorts of initiative (each accounting for about 6%) were the following:

- The subclass of class 1: Straight questions about material not currently in focus,
- The subclass of class 1: "I am not sure if <*stmt*>,"
- The subclass of class 3: The student does not understand what/when the tutor is talking about and is requesting repair, and
- The subclass of class 3: The student is not familiar with the physiology lingo.

Obviously, class 1 initiatives are by far the most common, and in fact Joel Michael and Allen Rovick, as the tutors, believe that many initiatives that I am classifying in other classes are really questions in "deep disguise." I do not really disagree with them, but I believe the distinctions I am drawing are important for the design of CircSim-Tutor.

It seems to us that the number and depth of initiatives rises as the student's grasp of the material rises, until at some point the student knows the material thoroughly and begins to simply answer questions, with few or no initiatives. In fact, in one session the student apparently had too little grasp of the material to be able to put together a coherent initiative. That student is one of the eight students who appear in two sessions, and in the second session the same student, having learned the material better by that time, generated six initiatives. That particular student is one illustration of this trend.

Responding to Student Initiatives

I would like to characterize the cognitive strategies that appear to determine how our expert human tutors respond to student initiatives. Below, I include interesting examples that illustrate these cognitive processes. Based on transcript analysis and discussions with our tutors, it appears to me that important aspects of the tutors' responses are the following:

- tactical use of hints,
- the tutors' concern with the students' use of correct physiological language, and
- factors which appear to affect the content, style, and expansiveness of the tutors' responses (examples occur about two pages below).

Our tutors have a definite preference for encouraging the students to solve problems for themselves. The tutors also focus attention on the students' problem solving skills.

The tutors are always willing to give careful explanations of how to reason, and they focus the students' attention on causal processes and causal reasoning. It appears that when the tutor believes a student is thinking causally and has a good grasp of the necessary information and reasoning skills, the tutors are willing to be more expansive and more general in responding to student initiatives.

As described above, I began investigating student initiatives by studying their occurrence and classifying them. The classification is oriented to the tasks of recognizing when a student initiative occurs and the student's intention. I then turned to studying the tutors' responses, hoping to solve a corresponding more difficult problem: understanding how our tutors respond to an initiative once it is recognized. The purpose of this investigation is to figure out how CircSim-Tutor should respond to student initiatives.

I have given particular attention to the responses by the tutors that seem to illuminate the tutors' cognitive processes when tutoring. In addition to generalizations about typical responses and about typical strategies, tactics, and goals, I will give examples of particularly interesting responses to initiatives. These example responses are not as typical as those I generalize about.

These sessions all occurred at the same point in the students' physiology coursework in their first year of medical school. Because our tutors are experienced Physiology professors, they know *what* material it is most important that the students learn from the sessions (i.e., what they need to learn at this point in their study of physiology), and they know what difficulties the students typically have at this point in their study. So, our tutors come to these sessions with pre-existing overall tutoring goals and considerable insight into the students' possible difficulties.

Nevertheless, the tutors had no advance knowledge about any specific student in these sessions. The tutors did not know how well any particular student in these sessions knew the necessary background facts, how well the student could apply that knowledge to make inferences, how skilled the student would be at thinking about the problems in the *right* way (i.e., causally), how well the students could explain their reasoning in the right language, or what misconceptions the student had. This situation matches the situation faced by CircSim-Tutor.

The most common initiative is a question about the topic that is being discussed/tutored at the point in the session at which the question occurs. The typical response to such a question about the locally current context is a brief answer, typically taking one or two sentences. At times, the tutor will defer an answer if the material will be covered shortly or simply ignore an initiative (especially questions that do not really make sense). If a question is not about the locally current context, especially if the question is more general and the answer would be more generally applicable, the tutors tend to respond expansively. See the example following the next paragraph.

In all these examples, the text that is the actual student initiative is shown underlined. Since we are now concerned with the tutor's response, the tutor turns in all the following examples are labeled with the tutor's initials rather than with "tu," even though I am not studying any differences between our two tutors. The student turns will continue to be labeled "st" (i.e., "student"). For the convenience of the reader, I have expanded all abbreviations that actually appear in the following examples and have corrected the spelling, as these changes do not affect the analysis being presented. In addition, at places I paraphrased a few words so the examples will be understandable without their surrounding context (these paraphrases are shown in italics inside square brackets *[like this]*). So, here's the example mentioned above that illustrates how the answer to a generalized question is fuller. To repeat, the actual initiative is underlined.

AAR: This is a tough one because you have: Cardiac Contractility decreased, Right Atrial Pressure increased, and Mean Arterial Pressure decreased. And you need to know which is the most powerful affect. It happens to be Right Atrial Pressure. So Stroke Volume is increased.

st: Ok, that was my problem. Will Right Atrial Pressure always be the major determinant of Stroke Volume?

AAR: Yes unless the primary change that occurs is an increase in Cardiac Contractility without a simultaneous change in the other determinants. Example, if something were to happen to the heart muscle to decrease Cardiac Contractility, Stroke Volume would fall. OK?

st: Yes.

[Session K31, turns 130–133]

Here is another example. In this case the student has just predicted that Stroke Volume is increased and takes the initiative by offering the rationale behind this prediction, implicitly asking the tutor to comment on the rationale. You will note that the tutor expands on this rationale, generalizing it. Note that the student's understanding and rationale, which triggered this expansion, was correct.

JAM: Do you want to look at your prediction for Stroke Volume again?

st: I thought it would increase because Right Atrial Pressure is the biggest determinant of Stroke Volume, and Right Atrial Pressure has increased.

JAM: You are correct, and this is a tricky one to predict for the following reason. The Direct Response change in Stroke Volume (due to the afterload changing) is small, and so we have to decide whether the Stroke Volume change in the Reflex Response is bigger than the Direct Response change or smaller. It can be argued either way. So your answer wasn't wrong. OK?

st: Yes.

[Session K37, turns 288–291]

While we have not yet carried out a quantitative analysis, it appears to me that there is a definite pattern to how expansively the tutors respond to students with different strengths. On the outer limits, it is clear that if the student is quite confused about how to reason through the problems then the tutor will focus sharply on presenting strictly causal reasoning involving only the seven physiological variables which the student is being asked to make predictions about. Such students never seem to draw more general answers from the tutors, and we believe this is a strategy of focusing the student's attention. It is also clear that if the student shows total mastery of the background knowledge and the skills of solving problems using that knowledge, the tutor will answer any question such a student does ask in a fairly extensive fashion. Having a strong grasp of the material, such students do not, however, normally generate many initiatives (they typically simply make all the right predictions and correctly answer all the tutor's questions). For students in between these extremes, there seems to be less clear cut evidence of a pattern to how expansively the tutors will respond. The tutors always have, in effect, prepared answers for questions from any student about *how to think* when solving the clinical problems: observation of the tutors reveals that they do not have to stop to create answers for this sort of question.

The following series of examples illustrates how the tutors seem to respond to students with a reasonable grasp of underlying facts and only mild confusion about how to solve the clinical problems. In some cases, there is an embedded tutoring dialogue with questions and answers rather than a straightforward answer to the original question. We are not sure just when or how the tutor will choose to do this, but we believe it may happen when the tutor needs further information about the student in order to respond optimally or when the tutor believes the student needs the answer in bite-sized pieces. The tutors definitely take things a step at a time when the student initiative (or the student's answer to a question from the tutor) reveals that the student is skipping over important intervening causal steps.

In the following example, the student offers an incorrect rationale for a previous incorrect answer (perhaps the student is still confused?). In response, the tutor makes a somewhat global meta-comment, along the lines of "here's how to think about this kind of problem."

> st: Stroke Volume *[is unchanged]*
> AAR: When Mean Arterial Pressure *[is]* increased, it's harder for the ventricle to pump blood. So what would that do to Stroke Volume?
> st: Stroke Volume would decrease . . . <u>I was thinking that Cardiac Contractility *[being unchanged]* would not allow more blood to be pumped out of the ventricle.</u>
> AAR: You need to take all of the determinants into consideration together.
> [Session K31, Turns 79–82]

In the next example we have another example of a meta-comment about how to reason; this excerpt is followed by a discussion (omitted here) of central venous pressure being determined by cardiac output, in which the tutor gives an expansive explanation of what happens and why.

> JAM: *[. . .]*, and if Cardiac Output is up what will change "next"?
> st: Stroke Volume
> JAM: Why would an increase in Cardiac Output cause a change to Stroke Volume?
> st: <u>Can I change that? Cardiac Output is determined by Stroke Volume, not the other way around. Cardiac Output determines Mean Arterial Pressure.</u>
> JAM: Exactly. It is much easier to make predictions about how the cardiovascular system will respond if you start with the perturbation (change) and reason causally (what determines what) from variable to variable. *[. . .]*
> [Session K48, Turns 64–68]

In the following example, by way of contrast, the student shows a strange misunderstanding, appearing to believe that the equation being discussed is *about* the Cardiac Output (thinking that it is determined by "pressures"). The tutor's response here is an example of a remedial didactic lecture for a rather confused student. Note the focused rather than expansive character of that lecture.

AAR: *[. . .]* Come back to Mean Arterial Pressure. What are its determinants?

st: <u>Mean Arterial Pressure depends on systole and diastole, however, I'm not seeing them directly in the *[Predictions Table]*.</u>

AAR: You are thinking of a way to calculate the approximate value of Mean Arterial Pressure. I'm thinking of a causal statement that says MAP = . Finish *[that equation]*.

st: Mean Arterial Pressure = pressure that

{THE TUTOR INTERRUPTS THE STUDENT'S INPUT AT THIS POINT}

AAR: Write an equation using only variables in the Predictions Table that says MAP = .

st: MAP = TPR X RAP ?

AAR: Close. MAP = TPR X CO. Remember?

st: Yes, Obviously the Cardiac Output includes those pressures.

AAR: I'd like to see you think about it differently. So let's see if we can get you to do so. Starting at the top, you correctly stated that the transfusion increased Right Atrial Pressure. Then you said that Stroke Volume went up. That's correct because Right Atrial Pressure is one of the important determinants of Stroke Volume. Well, when Stroke Volume went up, since the Heart Rate doesn't change in the Direct Response, Cardiac Output must have gone up (CO = HR X SV). Finally since Cardiac Output went up and Total Peripheral Resistance didn't change (it's a neurally controlled variable and we're in the Direct Response) Mean Arterial Pressure must have gone up (MAP = CO X TPR). That's the way to think about it. OK?

st: That is much clearer now.

AAR: You need to remember to think in causal sequences. *[. . .]*

[Session K39, Turns 70–80]

Turning to a different sort of tutoring tactic, the next example shows the tutor investigating a misconception from which the student does not suffer. It is probably helpful if we offer the reader some detailed comments on this excerpt, so I have included turn numbers for reference. The student has predicted earlier that heart rate and cardiac contractility are both decreased. If the student had responded to the tutor's initial question (in turn 102) by stating "decrease in sympathetic stimulation" (which occurs at turn 115), the tutor would have been happy. Instead, this strong student launches into an initiative which mentions *para*sympathetic stimulation, raising the tutor's suspicions. This first student initiative, the reader should note, is actually a question about the effect of decreased *sympathetic* stimulation on cardiac contractility. Interestingly, the tutor does not actually answer that question until turn 114, which is the end of the tutor's response to the initiative. The tutor's overall response to the initiative is to socratically cover, in detail, the effects of sympathetic and parasympathetic stimulation, first on heart rate, and then on cardiac contractility. Turns 105 and 106 are an embedded student initiative and response to clarify (repair) the tutor's question in turn 104. Apparently, since the student has just mentioned (in turn 103) that parasympathetic stimulation affects the heart rate, the tutor's question was taken to refer to something else.

JAM-102: OK, let's see what we have here. You started by predicting Heart Rate and Cardiac Contractility. Why?

st-103: Parasympathetic stimulation will decrease Heart Rate. With Cardiac Contractility, I know that sympathetic increases it, but will it be functionally decreased with lack of sympathetic stimulation?

JAM-104: We have to deal with two things here. First, what inputs to the heart determine Heart Rate?

st-105: What do you mean by inputs, SA and AV nodes?

JAM-106: No. *[The]* SA and AV nodes are structural pieces of the heart. What I mean is, what physiological signals reach the heart that determine the value of *[the]* Heart Rate?
st-107: Action potentials from the autonomic nervous system, either parasympathetic or sympathetic.
JAM-108: So, if [the] parasympathetic signal increases to [the] heart, what happens to [the] Heart Rate?
st-109: Decrease
JAM-110: And if [the] sympathetic signal to [the] heart decreases, what happens to [the] Heart Rate?
st-111: Decrease
JAM-112: Right. Think of parasympathetic and sympathetic as the brake and accelerator. Now, what inputs to the heart determine the value of Cardiac Contractility?
st-113: Sympathetic
JAM-114: Right. The parasympathetic portion of [the] autonomic nervous system has no functionally significant effect on Cardiac Contractility. So, why did you predict that Heart Rate and Cardiac Contractility would both decrease?
st-115: Decrease in sympathetic stimulation
JAM-116: Right. *[. . .]*
[Session K44, Turns 102–116]

In analyzing instances of hinting in our corpus of transcripts, Hume, Michael, Rovick, and Evens [1993] drew a clear distinction between hints which call the student's attention to some information by **p**ointing **t**o it (pt-hints) and hints which themselves **c**onvey **i**nformation (ci-hints). As you might expect, hints play an important role in responding to student initiatives. I now turn to some examples of these two kinds of hints. Here is a simple ci-hint.

AAR: It's the third effector that reacts. Which one is that?
[. . .]
st: <u>Volume is the other effector . . . The others are all affectors . . . Correct?</u>

AAR: The effector that contains alpha receptors is the vascular smooth muscle. Does that help?
st: Yes, *[. . .]*
[Session K31, Turns 24–31]

Now in the next example, showing a pt-hint, the student includes a quite incorrect rationale along with an answer, and the tutor responds by restating the question in more detail (implicitly a negative acknowledgement) so that the question directs the student's attention to the correct way of reasoning (pt-hint). The student in this example *already knows* the effect of the alpha-agonist.

AAR: *[. . .]* Now in what direction will Total Peripheral Resistance change?
st: <u>Total Peripheral Resistance will decrease because the blood will be pumped out of and back to the heart more rapidly.</u>
AAR: The question is, when arterioles constrict in response to the alpha agonist, in which direction will Total Peripheral Resistance change?
st: It will increase.
AAR: Correct. *[. . .]*
[Session K31, Turns 36–40]

A final category of particularly interesting tutor responses occurs when the student has problems making a statement in the correct physiological language. Teaching the language of physiology is one of the explicit goals of our tutors. In fact, it is this concern with language that initially prompted the CircSim-Tutor project. CircSim-Tutor must have sophisticated language abilities if it is to effectively help students learn the language of physiology.

In the following example, the student is actually confused about language, and the tutor picks up on this problem. Note that the tutor goes on to ask a question about the definition of another technical term, which is type of question that does not occur frequently. In this example, as is typical, tutoring about language is not completely separable from tutoring about content. We consider that to be an important generalization about the tutoring sessions we have analyzed.

st: I guess I'm confused because doesn't vasoconstriction decrease the blood reservoirs in veins, which increases venous return?

JAM: Well, when we talk about vasoconstriction we are almost always talking about changing the size of arterioles. If we want to talk about changes to the veins (capacitance vessels) we usually talk about VENOCONSTRICTION. The key idea to be pursued here is afterload. Do you remember what this means?

[Session K37, Turns 209–210]

Conclusion

In this chapter I have presented an analysis of interesting examples which illustrate the cognitive processes of expert human tutors in responding to student initiatives. Our tutors have a definite preference for encouraging the students to solve the problems for themselves, and the tutors focus attention on the students' problem solving skills. Despite the tutors' preference for hinting rather than simply answering questions, the tutors give careful explanations of how to reason whenever requested. In addition, the tutors focus the student's attention on causal processes and causal reasoning. It appears that when the tutors believe the student has a firm grasp of the need to reason causally and a good grasp of the information and reasoning skills needed to solve the clinical problems covered in these sessions, the tutors are willing to be more expansive and more general in responding to student initiatives.

CHAPTER IV

ARCHITECTURAL INTEGRATION OF DISCOURSE GENERATION IN CIRCSIM-TUTOR

The overall architectural plan for how the Discourse Generator should be integrated into the CircSim-Tutor system is in a state of great ferment, as new people have joined the project and are exploring many alternative architectures in ongoing discussions. I believe this is beneficial, but in order for me to present a plan here for generating larger-scale discourse and dialogue structures I must assume some fixed architectural context. The approach I present here reflects the thinking that was current during the summer of 1994 (which, in my opinion, is the last time there was a coherent overall architecture for integrating these sorts of extended discourse structures into CircSim-Tutor). That approach, presented throughout this chapter, was something of a waterfall model in which the Instructional Planner was supposed to feed the Discourse Generator, and the Discourse Generator was supposed to feed a Sentence Generator. In contrast, our current thinking is that instructional planning and discourse planning need to be done by one *integrated* planner rather than as two sequential stages in generation. This is a point of view I have advocated over the past three years; however as a group we have not yet reached agreement among everyone concerned about how to structure the CircSim-Tutor system using that kind of integrated approach.

As seen from the Instructional Planner and the rest of the system, the Discourse Generator and the Sentence Generator, together, were to look like a single integrated module known as the Text Generator. What this really means is that communication with the Text Generator is really communication with the Discourse Generator. Inside the Text Generator, the Discourse Generator feeds the Sentence Generator. The text being generated basically appears as a side-effect when the Sentence Generator runs (the text is fed to the Screen Manager). The Sentence Generator returns some sort of success/failure information to the Discourse Generator, indicating whether or not the Sentence Generator succeeded in producing text. The Discourse Generator then returns this success/failure status as the return value from the Text Generator as a whole.

		Text Generator				
Instructional Planner	$\leftrightarrow$	Discourse Generator	$\leftrightarrow$	Sentence Generator	$\leftrightarrow$	Screen Manager

The basic proposal for the interface between the Instructional Planner and the Text/Discourse Generator is that there will be a well-defined formal representation for what the Instructional Planner wants the Text Generator to generate. We call this formal representation a "Logic Form."

Semantic Forms

Corresponding to Logic Forms, we have a formal representation used to communicate from the Discourse Generator to the Sentence Generator. That representation is called a "Semantic Form." We have a few different kinds of Semantic Forms, but the most common type, which I will call the generic type, translates into an independent clause or a sentence. This generic kind of Semantic Form incorporates a Logic Form as its "Pred" field (Pred stands for Predicate). This makes sense if we keep in mind that the Discourse Generator is really elaborating requests from the Instructional Planner. In the process, the Discourse Generator performs the following key tasks.

- Choose the tense — normally present tense.
- Track the Discourse Focus and pass it on to the Sentence Generator. In Semantic Forms, the Discourse Focus field is named "DFocus."
- Choose the Mood (Declarative, Question, Imperative, Let's) and include it in the Semantic Forms. The "Let's" mood is used for sentences like "Let's look at your predictions." or "Let's think about the heart rate."

When the discourse focus changes, this is communicated in advance by a type of Semantic Form that does nothing else. The generic semantic form, which translates into an independent clause or sentence, has fields for Mood, Tense, Topic and Pred.

To give an actual example, suppose the Instructional Planner passes the following Logic Form to the Discourse Generator.

```
( Statement ( Determines   ( )
              ANS   ( )
              IS  ( )  ))
```

This Logic Form represents a declarative statement that says, "The ANS determines IS." (IS means Inotropic State.)

The corresponding yes/no question, "Does the ANS determine IS?" would be represented by the following Logic Form.

```
( Y/N-Question ( Determines   ( )
                 ANS   ( )
                 IS  ( )  ))
```

Questions other than yes/no questions will have a missing constituent, and the missing constituent is what is being asked about. We can see that by comparing the Logic Forms given above with the following Logic Form which means, "What determines IS?"

```
( Question ( Determines   ( )
                          ?
                          IS  ( ) ))
```

So, the examples above represent Logic Forms that the Instructional Planner could pass the Discourse Generator. Suppose the last of the three Logic Forms given above,

```
( Question ( Determines  ?  IS ) )
```

were to be passed to the Discourse Generator. If doing so were the very beginning of the dialogue, we would start out with the Discourse Focus being unknown. Thus the Discourse Generator would start out by passing the following Semantic Form to the Sentence Generator.

```
( DFocus  !Unknown!  (  ) )
```

!Unknown! is a global constant, intended to be used the way it is used in the example above. The empty set of parentheses at the end of the form is the "modifier slot," and I will defer explanation of modifier slots other than to note that every "thing" appearing in a Semantic Form will have one. After starting out with the above Semantic Form, the Discourse Generator would next pass the Sentence Generator the following semantic form, representing the question itself.

```
( ( Topic  DFocus
  ( Mood  Question )
  ( Tense  Present )
  ( Pred  ( Determines (  )
            ?
            (IS  (  ))  ))  )
```

In this Semantic Form, modifier slots occur after the word "Determines" and after "IS." Taking the Semantic Form one piece at a time, the first line says that the Topic of the sentence is the DFocus, which we already said is unknown. Obviously, we could literally repeat the value of DFocus. Instead, I have chosen to treat DFocus and Topic as variables that can appear in place of their values. Typically, the Topic will be the same as the DFocus, indicated as in the example above. One can also use the Topic variable in the Pred field. I have done this so that the Sentence Generator will not have to do anything to notice that some form is actually the current value of the DFocus or of the Topic, thus sparing that module from unnecessary comparisons. Correspondingly, the Sentence Generator will need to have actual variables for DFocus and Topic, will need to save the actual values in these variables, and will need to use the actual value (substitute the actual value) everyplace "DFocus" or "Topic" appears in a Semantic Form. The next two pieces of the Semantic Form above say that it is a question and should be in present tense. The predicate is the same as the Logic Form we began with, simply reformatted here for readability. Notice that this whole Semantic Form has a set of parentheses wrapped around it. Translating this Semantic Form into the input format for existing sentence generators such as FUF turns out to be fairly straightforward.

To illustrate the role of the modifier slots, let's consider the following sentence: "You have is slightly, but importantly, wrong." The reader can see that the underlying core of the statement is, "You have it wrong." This is elaborated in two ways: first that the degree is slight and second that the importance is high. We can show this with the following RST-style diagram, Figure 4.

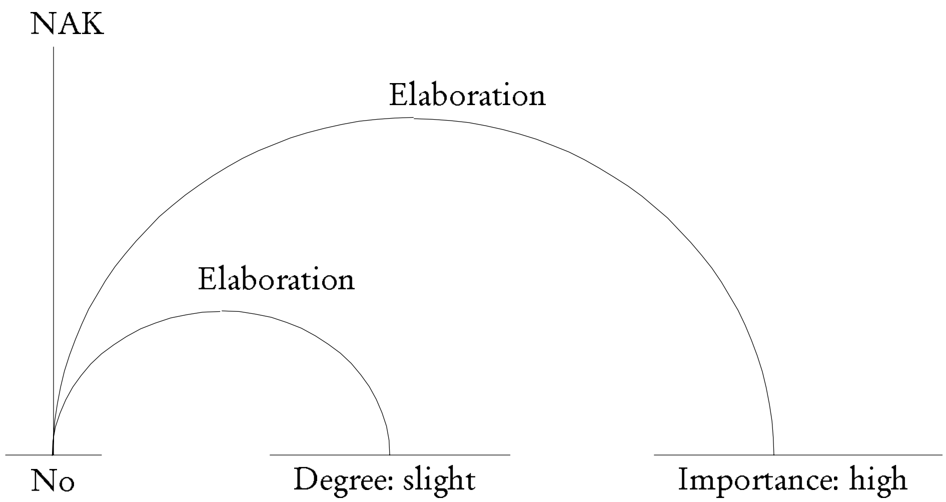

Figure 4. RST-style Diagram For, "You have it slightly, but importantly, wrong."

If we consider appropriate Semantic Forms for this sentence, they would look like this.

```
( DFocus  <previous student answer> )

( ( Topic  DFocus
  ( Mood  Declarative)
  ( Tense  Present)
  ( Pred  ( BE   (  )
          Topic  (  )
          Incorrect  (AND
                        ( Degree   slight )
                        ( Importance   high )  )  ) )  )
```

In the above example, we have two Semantic Forms. The first telegraphs that the student's previous answer is becoming the Discourse Focus. The second represents the sentence, "You have it slightly, but importantly, wrong." Notice that the modifiers about degree and importance apply to the wrongness of the answer. Thus, in the "Pred" of the Semantic Form, these two modifiers go in the modifier slot for "Incorrect" but "BE" and "Topic" have no modifiers.

We had some discussion among members of the CircSim-Tutor project about whether or not these ubiquitous modifier slots are really necessary. In fact, the syntax of English allows *any* constituent of a sentence to be modified (e.g., by adjectives, adverbs, prepositional phrases and so forth). In light of this, I have chosen to include a modifier slot for each "thing" that occurs in a Semantic Form. Regularizing the Semantic Forms in this respect will make them easier for programmers to deal with, since their structure will have no special cases.

The CircSim-Tutor system has a discourse log file that contains, among other things, all the Semantic Forms passed to the Sentence Generator. The Discourse Generator is responsible for adding them to this log file. In addition, the log file may eventually contain various sorts of information about the overall intentions behind what the Discourse Generator does, but this aspect of planning intentions in discourse generation has not yet been worked out.

CHAPTER V

GENERATION OF EXPLANATIONS, SUMMARIES, AND MULTI-TURN STRUCTURES IN TUTORIAL DIALOGUE

In this chapter, I explain how we can generate larger-scale discourse and dialogue structures in tutorial dialogue. Generating multi-turn structures such as a directed-line-of-reasoning, as well as generating extended single-turn structures such as explanations and summaries, requires interaction of many parts of the CircSim-Tutor system. I will begin by discussing how to generate a summary of the correct solution of the Direct Response phase of the clinical problems presented to students in CircSim-Tutor. Thereafter, I will explain how the same techniques can be used to generate summaries of pieces of the solution. Finally, I will define "directed line of reasoning (DLR)" structures, and explain how they may be generated.

Summaries of Partial or Complete Solutions

In the keyboard-to-keyboard tutoring experiments with human tutors, many students suggested that CircSim-Tutor should always summarize the correct solution of the Direct Response phase of each clinical problem before moving on to tutoring about the Reflex Response phase. The students correctly observe that it is possible for them to solve the Direct Response phase correctly without being sure that their reasoning is correct.

This presents an interesting problem since most clinical problems begin with some material (such as artificial pacemakers, human centrifuges, and so forth) that is not in the knowledge base in CircSim-Tutor. As a result, part of the explanation really needs to be written by human experts and stored as canned text. Specifically, the part that starts from the clinical scenario presented in the problem and goes up to the first variable to be affected in the predictions table [Rovick and Michael, 1992]. Here is an example of this initial portion of an explanation that would need to be stored as canned text.

> The centrifuge displaces blood to the legs and lower abdomen (venous capacitance), reducing venous return to the heart. The result is a reduction in the central blood volume (CBV), and this reduces Central Venous Pressure (CVP). Thus, the first variable that is affected in the predictions table is CVP.

Once we get to the first variable to be affected in the predictions table, we have the necessary information in the knowledge base of CircSim-Tutor to support the automatic generation of a high-quality summary of the rest of the Direct Response phase.

The next step needed when generating this summary is to obtain a solution from the CircSim-Tutor knowledge base. Doing so requires a certain amount of arbitrary but well-defined software magic, and the knowledge base provides functions that return such a solution. I will not present the details of these calls here, since they are not very readable. The output, however, is reasonably straightforward. In the case of the pacemaker problem mentioned earlier (heart rate determined solely by an artificial pacemaker, which suddenly fails, significantly increasing the heart rate) the solution output for the Direct Response phase is:

```
'( ( *HR*  increase )
   ( *HR*  increase  *CO*  increase )
   ( *CO*  increase  !!MEAN-ARTERIAL-PRESSURE!!  increase )
   ( *CO*  increase  *CVP*  decrease )
   ( *CVP*  decrease  *SV*  decrease ) )
```

What the above output really means is the following. Heart Rate (HR) increases. The increase in HR causes the Cardiac Output (CO) to increase. The increased CO causes Mean Arterial Pressure to rise. The increased CO also causes the Central Venous Pressure

(CVP) to fall. The decrease in CVP causes Stroke Volume (SV) to decrease. That's the end of the direct response phase.

In order to generate the proper input for a Sentence Generator, we need to note that the first step in the solution is special since it is more or less the direct effect of the clinical problem. Similarly, the final step is special since it is the end of the solution and should be identified as such in the surface text that is generated; the last step can be preceded by "Finally," or it can be followed by a sentence that is like the last sentence in the preceding paragraph. The Discourse Generator can point out that the increase in CO has two effects (increasing MAP and decreasing CVP). In addition, the Surface Generator needs to be told that the discourse focus shifts with each step, so that it can choose to pronominalize each previous step. If the Discourse Generator does all these things, CircSim-Tutor might plausibly generate a summary like the following (omitting the description of the failing pacemaker, which I have said would be canned text).

> Let's summarize the Direct Response. The pacemaker fails, causing HR to increase significantly. The increased HR causes CO to rise. This increases MAP and, in addition, causes CVP to fall. Because CVP falls, SV also falls. And that's the end of the Direct Response phase.

I have actually implemented the generation of the necessary Semantic Forms to allow this. However, our physiology experts have not yet had time to write the canned text part of the explanation for all the clinical problems (For example, "The pacemaker fails, causing HR to increase significantly."). The Discourse Generator can easily identify the first and last steps in the solution. So, we begin by generating Semantic Forms to introduce the summary.

```
((DFocus (DR ( )))

((Topic DFocus)
 (Mood Lets)
 (Tense Present)
 (Pred (Summarize ( ) Topic)) )
```

At this point, the Discourse Generator should tell the Sentence Generator to spit out the canned text for the clinical problem. Since that text does not exist, for the time being I need to set the Discourse Focus to be the first variable that will be affected in the predictions table. We get this from the first line of the solution returned by the problem solver code, which for the pacemaker procedure is (*HR* increase). Accordingly, the Discourse Generator will put this out as a Semantic Form.

```
(DFocus !!Heart-Rate!!)
```

Next, the Discourse Generator has to tell the Sentence Generator to say that it is the first variable to be affected. In the CircSim-Tutor project, we have been calling this the "primary variable," and the student's problem solving task really starts from the primary variable. I generate a Semantic Form identifying it.

```
((Topic (Primary-Variable ( ))
 (Mood Declarative)
 (Tense Present)
 (Pred (Primary-Variable ( ) (!!Heart-Rate!! ( )))) )
```

From this point, I simply traverse the rest of the list of steps in the solution, generating Semantic Forms for each step. These Semantic Forms basically say that the previous step causes a change in whatever variable it affects, along with the direction of the change. If two variables are affected, as is the case for CO in the pacemaker procedure, I generate an "ADDITIONALLY" Semantic Form between the two. It appears that this approach will work for *any* clinical problem we might want to concoct. Somewhere in the sequence of cause-and-effect steps, MAP will be affected. When this happens, I remember its direction of change for later reference.

At the end of the solution path, I put out three Semantic Forms that wrap up the summary. The first of these basically means, "That's all she wrote." The second say that, "Because this is DR, the neural variables are unchanged." The third is a reminder, "Note that at the end of DR, MAP is <*value*>." (where <*value*> is the direction of change in MAP, increased or decreased, that was remembered from the solution for later reference here). Because our human tutors normally leave the initiative with the student at this point, in effect offering to explain further if the student wishes, I end up by putting out one more Semantic Form that translates into, "OK?"

If we decide that a sub-part of a solution should be summarized, we can use the same techniques. We simply get the solution from the problem solver, pick out the subset of the solution that is to be put out as a summary, and then go through that subset, a step at a time, using exactly the same techniques as for a summary of the entire Direct Response phase. This sort of summary of part of a solution is appropriate when that subpart of the solution has required extended tutoring with the current student. Our human tutors do this. We have had some discussions over how to decide when these summaries are needed. I believe the Discourse Generator should simply track how many turns have been required to get from one step to the next. When the discussion moves to a new topic, the Discourse Generator should produce a brief summary of the piece of the solution that has involved extensive tutoring. Some members of the development team feel this is a decision that should be made by the Instructional Planner. We have not tried to implement such summaries as yet.

Directed Line of Reasoning (DLR)

A Directed Line of Reasoning or DLR is a series of bite-sized leading questions and answers that are intended to evoke correct cause-and-effect reasoning from the student, based on information the student already knows. DLRs thus appear when the material in question is a chain of causal steps, and when the student plausibly already knows all the steps. DLRs play various roles in the tutoring sessions with our human tutors. They may serve as hints, where the tutor leads the student toward, or even to, something the student did not manage to produce without help. They may serve as a summary, allowing the tutor to verify that the student really knows all the steps, and in the process reviewing the sequence of the steps for the student. They can even serve as an explanation for a student who "almost" knows the content of the desired explanation. Here is an example of a DLR (adapted from session K12, beginning with turn 65). It begins and ends with an explanation, which is a typical feature of DLRs. For the convenience of the reader, I have expanded the abbreviations in square brackets.

tu: Since we are now in the Reflex Response period, the variables that change first are the ones that are neurally controlled. Which of these variables would be affected first?
st: CC [Cardiac Contractility]
tu: Of course! And in what direction?
st: Decrease
tu: Right again. And how would that affect SV? [Stroke Volume]
st: Decrease
tu: Sure. And what effect would that have?
st: Decrease CO [Cardiac Output]
tu: Yes again. Then what?
st: MAP d [Mean Arterial Pressure decreases]
tu: Yes again. And it is MAP that is regulated by the BAROceptor reflex, which is why it is called that.

Notice that one role of the explanation at the end is to let the student know that the DLR is over, and note that all the tutor productions in the middle of the DLR let the student know that there is still another step. Also, you will notice that on a surface level there are some characteristic ways that language is used in a DLR, such as the ongoing parallel use of "And" to introduce each new question in the DLR. This is part of how tutors seem to flag the continuation of the DLR, thus letting the student know that what is expected is a bite-sized answer. If the Sentence Generator is going to generate these sorts of DLR language, it must be informed that a DLR is underway.

The domain knowledge that the Discourse Generator needs in order to generate a DLR is fairly similar to the domain knowledge used in generating a summary of the Direct Response. In particular, we must get a cause-and-effect chain. This chain is normally at a given level of the knowledge base, and typically is a series of steps involving variables in the predictions table. The series of steps in the summary of the Direct Response used as an example above is typical of such a causal chain. Each bite-sized piece is one step in that chain (one line in the solution from the knowledge base).

Since the Instructional Planner actually decides when we should attempt a DLR, the Instructional Planner presumably has already gotten that chain from the knowledge base, and therefore the Discourse Generator can get it as well (whether the Instructional Planner passes the chain as part of the Logic Form or the Discourse Generator gets it directly from the knowledge base). I will say nothing further about this aspect of generating a DLR. The Instructional Planner will tell the Discourse Generator to do a DLR and will say what material it should cover (start with this variable and end with this one) and at what level of detail. The level of detail may be chosen to correspond to what the student model says the student most probably knows, but this is really not a discourse planning issue per se. The crucial question during a DLR is whether or not the student is managing to take each bite-sized step by answering the bite-sized questions correctly. Let's take up that topic.

The current thinking about how the student's productions will be processed says that they are parsed and then passed to a "Judger" routine that determines whether or not the student's statement is correct. In my opinion, the Discourse Generator should make the expected answer available so that the Judger can decide whether or not the student's answer is what the Discourse Generator expected. This is controversial; other people in the project believe the questions have self-evident correct or incorrect answers and that the Judger needs no help from the Discourse Generator to decide whether the student's answer is as expected. However this judgment process happens, the student's answer is eventually evaluated as correct or incorrect. If the answer is correct, the DLR can continue. If not, or if the student takes the initiative (e.g., by asking a question) then the Instructional Planner needs to replan.

Since the Instructional Planner normally does something at each turn, its role changes during a DLR. Our thinking is that once the DLR begins, the Instructional Planner should simply observe and tell the Discourse Planner to continue with the DLR. The Instructional Planner can see what the Discourse Planner is doing (or has done) by observing the log file.

The Discourse Planner will start the DLR by telling the Sentence Generator that a DLR is beginning. There is a reserved Semantic Form for the purpose. When the DLR is over, the Discourse Generator will also inform the Sentence Generator of that fact, again by using a reserved Semantic Form. In between the beginning and end of the DLR, the communication between the Discourse Generator and the Sentence Generator is no different than usual. Each basic turn in the middle of a DLR consists of a positive acknowledgment of the preceding correct answer and then a new bite-sized question.

Content-specific Explanations

Earlier, I discussed the use of schemas as a means of generating rhetorical structure, and I criticized schemas as being too inflexible for that purpose. Schemas can play a different role in generating explanations, however. At times, experience has shown that the optimal *content* of a specific explanation would not fall out from more generally applicable rules. One such example occurs in explaining the difference between the Frank-Starling effect vs. Inotropic State (IS) of the myocardium. Explanations on this topic may be generated from a schema in order to generate an explanation with the optimal content.

Let me explain the Frank-Starling effect here. Systole is the period of time when the heart is contracting and diastole is the period of time when the heart is relaxed between contractions. The Frank-Starling effect is a non-linear length-tension relationship of muscle fibers. A relatively small increase in the muscle fiber tension that is present at the beginning of contraction will cause a relatively large increase in the contractile force developed. Now as applied to the heart, the tension of the muscle fibers is a function of the volume of blood in the heart at the beginning of a contraction, which we call end diastolic volume (EDV) or filling. If we draw a graph showing EDV as the independent variable on the horizontal axis and showing SV as the dependent variable on the vertical axis, we get an S-shaped curve showing a direct relationship; increased EDV causes increased SV. The practical effect can be described in a quite different way as the Frank-Starling Law of the Heart: "During systole, the heart pumps out the volume of blood that came into it during the preceding diastole. In other words, the amount of blood left in the heart at the end of systole is more or less a constant."

Students frequently seem confused about IS, typically demonstrating a belief that a change in filling volume (end diastolic volume) will affect IS. This confusion may be even more pronounced when the term "Cardiac Contractility (CC)" is used instead of IS. Thus, our human tutors actively look for this particular confusion and it must be remediated relatively often. As a result, they have discovered through long experience that the best didactic explanation has the following structure.

- Say, "You are confusing IS with the Frank-Starling effect."
- Say, "They are not the same."
- Define the Frank-Starling law, ending up by saying that the stroke volume is a function of the end diastolic volume.
- Say, "In contrast."
- Define IS, including the fact that it is controlled by the autonomic nervous system, and wind up the definition by saying that, "If IS increases (positive inotropic effect), then stroke volume increases with end diastolic volume (filling) held constant."
- Finally, state that, "An increase in IS shifts the Frank-Starling curve upward and to the left."

Note that the last point, about a change in IS effectively shifting the Frank-Starling curve, is *not* something that would ever get generated from general principles. The desired explanation is somewhat as follows.

> You are confusing the Frank-Starling effect with IS. They are not the same. You will recall that the Frank-Starling effect is a length-tension relationship of the muscle fibers in the heart. An increase in filling or preload (EDV) results in an increase in SV, and vice-versa. The Frank-Starling law means that SV changes when EDV changes.
>
> In contrast, IS is determined by the autonomic nervous system. A change in IS will cause a change in SV with EDV held constant. In effect, an increase in IS (positive inotropic effect) will shift the Frank-Starling curve upward and to the left.

Choice of How Interactive to Be

As has been mentioned, the tutor must choose how interactive to be. These choices are based on what the student appears to know and how well the student is doing at applying this knowledge. The tutor has various choices available. At the less interactive end of the scale, the tutor can respond to a question by simply giving the answer. Similarly, the tutor can explain something by, well, simply explaining it as a little didactic lecture. At the other end of the scale, the tutor can give maximally vague hints.

Remediation of any topic can be broken into units, and each unit can be covered by simply telling the student, by trying a DLR, by giving a ci-hint, or by giving a pt-hint (listing the options ordered by increasing interactivity). In giving a summary, the tutor can similarly simply give a little didactic lecture that simply explains or summarizes or can alternatively try a DLR.

In a sense, the tutor's choices can be thought of as vaguely like a two-dimensional matrix. We can think of the size of the task on one axis (e.g., a summary, occurring after remediations, that each consist of various explanations, with each explanation containing various units). On the other axis, we can think of the degree of interactivity. However, even though we can list about four different choices for interactivity when covering a unit, there are no real corresponding four choices for each explanation, remediation, or summary. Thus, this matrix metaphor should not be taken literally.

CHAPTER VI

CONCLUSION

Summary

In this thesis, I have presented studies of student initiatives in keyboard-to-keyboard tutoring dialogue, and the tutors' responses to them. I also explained how we can generate extended summaries, explanations, and directed-line-of-reasoning examples like those produced by our human tutors, Joel Michael and Allen Rovick.

In reviewing the literature, I have identified many issues that we can consider in designing and improving CircSim-Tutor. Particularly significant is the suggestion that we may enhance student understanding and recall of material if we try to have CircSim-Tutor take a consistent point-of-view in its explanations and summaries, as well as in its question-and-answer exchanges with the students. Along these same lines, the work of Wick and Thompson [1992] suggests, in my opinion, that appropriate responses to student initiatives will (or should) involve a conscious choice of a "line of explanation," and that this *may be different from* the "line of reasoning" actually used by the student or tutor in solving a problem.

The work of Graesser and Hemphill [1991] suggests that CircSim-Tutor should make an obvious point of telling students to reason in terms of physical cause-and-effect, not in teleological terms of purpose. Their work also shows that this is not the usual style of reasoning in biology, so that CircSim-Tutor should maintain a "high index of suspicion" that students will have a teleological reasoning style. Our keyboard-to-keyboard transcripts contain clear examples where our tutors make a point of this with some students and where those students appear to be helped significantly by this strategy. Graesser and Franklin [1990] also gave factors that differentiate good and poor answers, which we could easily incorporate into rules for CircSim-Tutor to use in answering students' questions.

With respect to the student initiatives in our keyboard-to-keyboard transcripts, I have given a classification scheme for them, described various studies of inter-rater agreement using that scheme, and discussed the results of those studies. In the opinion of all six persons who have used it to classify the initiatives, the classification scheme does describe the initiatives occurring in a substantial body (about 58 hours worth) of keyboard-to-keyboard tutoring transcripts. Unfortunately, our tutors (Joel Michael and Allen Rovick) have not yet had time to classify the initiatives using my classification scheme; their input would be very valuable. I also studied the tutors' responses to the student initiatives, paying particular attention to the factors that appear to affect how the tutors respond.

I have developed algorithms to generate summaries, explanations and multi-turn structures such as Directed Line of Reasoning (DLR) exchanges. I have shown that important commonalities underlie all these tutorial discourse structures, particularly with respect to the knowledge that the tutors must consult. Little or no previous work that we are aware of has been done on generating multi-turn structures such as DLRs. I have discovered that in our keyboard-to-keyboard transcripts DLRs serve various roles in tutoring: as summaries, as extended hints, and as a form of explaining a chain of cause and effect. I have also explained how to integrate them into a proposed overall architecture of CircSim-Tutor.

Future Work

I have not yet had time to do any significant work on exploiting Johanna Moore's [1995] work appropriately. To some degree, this reflects the fact that Moore was discussing the generation of extended turns in advisory dialogue, where we are concerned with generating rhetorical structures for a much more interactive dialogue. It also reflects the fact that so far I have been working mostly on getting discourse and dialogue planning integrated into the CircSim-Tutor system and on the generation of summaries and DLRs. Moore accomplished rule-based generation of text having an RST structure and showed that the texts satisfied several interacting constraints in a reasonably optimal fashion. The fact that Moore's system is rule-based and my Discourse Generator is more procedurally based has slowed our progress in adapting her work. As we move to a more rule-based approach, especially integrating instructional planning with discourse planning, I believe Moore's work will be of significant use to us. The fact that I have not yet been able to make appropriate use of her work is perhaps the most disappointing aspect of my progress so far. I want to work on incorporating her work now that we are actually moving to a rule-based approach.

Moore showed that a discourse generator must know and record the intention behind everything that it generates, in order to handle follow-up questions. For this reason, we need a largely intention-based discourse planner. On the other hand, we need some schemas in order to capture the content of certain explanations such as the one contrasting the Frank-Starling effect vs. Inotropic State (the explanation must end by saying that increasing IS is like shifting the Frank-Starling curve upward and to the left). Thus, our planner needs to be a mix of intention-based and schema-based.

In human tutors, instructional planning and discourse planning are not really separate processes, but rather each deeply influences the other. I feel strongly that CircSim-Tutor would produce better discourse if it also integrated its instructional and discourse planning. As near as I can tell, this planner would be rule-based (declarative knowledge) and would probably choose its rules by a process of unification. Recently, others in the project seem to have come around to my point of view about the need for integrating these two types of planning, but I have not had time to do any significant work on this so far since my work on the previous waterfall-style model was so far advanced. Backtracking and recasting my work into an integrated planner appears important. So, I would like to recast the current discourse generation, which has a procedural flavor, to a set of rules suitable for use with an actual planner or planning engine, should a persuasively firm choice of a specific actual planner ever materialize. However, the instructional planning done by Ramzan Ali Khuwaja must be re-engineered as well, and that is apparently a quite large task. Reva Freedman has already done significant work on this task.

There are a couple of smaller items for future work. I would like to incorporate the work of Graesser and Franklin [1990] mentioned above. I could generate something closer to the input format for a specific sentence generator such as FUF.

I had proposed possible extensions or reworking of RST for tutorial dialogue. As work progressed, it became clear that while minor extensions such as adding a relation for a *chain* of cause-and-effect would help, what really appears necessary is some sort of grand theoretical scheme of discourse and dialogue structure. Some vague possibilities are taking Hans Kamp's Discourse Representation Theory (DRT) into account, or perhaps some approach that is partly unification-based.

The work by Moore and Pollack [1992] showing that the actual physical structure of an RST analysis is unsatisfactory has led me to believe something more radical than minor extensions and reworking of RST is called for. In my opinion, you can easily show that speakers, especially tutors, have goals going on that cannot necessarily be reverse-engineered from the dialogue alone and which may not result in any surface realization for an extended period of time. So, in short, I think any suitable model must include multiple levels of analysis and not generate just a single hierarchical structure.

Given a fairly complete version 3 of CircSim-Tutor, it would be quite interesting to try retargeting the working version 3 of CircSim-Tutor as a remedial mathematics tutor for first-year college students. Interesting questions are: what would need to be added, removed, changed in order to make the system usable for such a different domain, and what would be (or could be) invariant?

A great deal of further work on analyzing the tutors' responses to initiatives is possible. Here's a list of questions that I think are worth investigating. Some of these may be of interest to Farhana Shah, a new member of the CircSim-Tutor development team. I intend to investigate many of these questions in the immediate future.

* How often does the tutor answer a student's initial question with a question?
* How often does the tutor answer a question with a question in extensions of the exchange that develops in response to an initiative?
* How often does the tutor respond to a question by requesting information from the student?
* How often does the tutor request information during extensions of an exchange that develops in response to a student initiative?
* How often does the tutor respond to a question by making some sort of an offer (for example, by offering to explain something)?
* How often does the tutor make some sort of offer during extensions of an exchange that develops in response to a student initiative?
* How often does any reasonable equivalent of the tutor's response require reference to (i.e., knowledge of) the student's prediction(s)?

* How often does any reasonable equivalent require reference to/knowledge of the preceding dialogue context?
* How often does any reasonable equivalent require reference to/knowledge of both the preceding dialogue and also the student's predictions?
* How often is the response crucially based on the user model?
* How often is it based on a user stereotype rather than a model of the particular student?
* How often is the response plausibly user-independent (e.g., a factual answer)?
*How often do tutors assume ***erroneously*** that the student knows something (post-hoc analysis, obviously)? This appears quite important to study.
* How often does it appear the student is stuck when he/she generates an initiative?
* How often does the tutor provide a plan, and how often a hint at a plan?
* How often does the tutor provide a general overall goal or frame of thinking, and how often a hint at a goal or frame?
* How often does the tutor give one step, and how often a hint at a step?
* Do student's seem forthcoming about their difficulties, and in particular do they communicate them cooperatively on request?
* Do students follow up if not satisfied? (Can we even study this?)
* How often do students appear satisfied with the response from the tutor?
* How often are students apparently able to use the tutor's response effectively?
* *When* do they use the tutor's response effectively, and when not?
* If they don't use it effectively, how often does the tutor try again *on the same line as the response that failed* ?

* How often does the tutor answer only *part* of the student's question (giving just a next step, or whatever)? We expect this normally happens when doing so is a more helpful response than a full answer would be.
* How often does the tutor not answer the student's literal question at all, because the questions reveals (or is) an incorrect plan?
* When asked such a question, how often does the tutor warn the student that the student's plan appears to be incorrect?
* When asked such a question, how often does the tutor actually correct the student's plan?
* How often does the tutor answer a different question than was asked (including in response to questions based on incorrect plans)?
* What fraction of student initiatives reveal student misconceptions?
* When this happens, how often does the tutor intentionally *not* immediately correct the misconception?
* On the basis of inferred or known high-level goals of the student, how often does the tutor provide information that was not asked for but which is more relevant than the information that the student asked for?
* What sort of language or discourse tactics do the tutors use to "brush off" a student initiative (whether to defer it or to ignore it)?

Many of the above questions are stated as "how often" since that seems fairly easy to investigate. In the case of those aspects that turn out to be frequent, we could also try to investigate when (and why) they occur. This will probably be more difficult than merely determining their frequency of occurrence.

In summary, the task listed above that I am most interested in is my continuing attempt to come up with some sort of grand theoretical description of discourse and dialogue structure. I am also interested in the issues involved in retargeting CircSim-Tutor as a remedial college mathematics tutor, since I am involved in teaching college students who are weak (and lack confidence) in mathematics. As a result, I am quite interested in understanding what kind of help such students would find most effective and inspiring.

BIBLIOGRAPHY

Abelson, Robert P. (1975a). "The Reasoner and the Inferencer Don't Talk Much to Each Other." In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1 (Supplement).* Morristown, NJ: ACL. pp. 3–7 of the Supplement.

Abelson, Robert P. (1975b). "Does a Story Understander Need a Point of View?" In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1.* Morristown, NJ: ACL. pp. 140–143.

Allen, James F. (1983a). "Maintaining Knowledge About Temporal Intervals." *Communications of the ACM*, 26(11), pp. 832–843. (Reprinted in Brachman, Ronald, and Levesque, Hector, eds., 1985. *Readings in Knowledge Representation.* Los Altos, CA: Morgan Kaufmann. pp. 510–521.)

Allen, James F. (1983b). "Recognizing Intentions from Natural Language Utterances." In Brady, Michael, and Berwick, Robert C., eds., 1983. *Computational Models of Discourse.* Cambridge, MA: MIT Press. pp. 107–166.

Appelt, Douglas E. (1985a). "Planning Natural Language Referring Expressions." *Artificial Intelligence,* (25)1, pp. 1–33. (Reprinted in McDonald, David D. and Bolc, Leonard, eds., 1988. *Natural Language Generation Systems.* New York: Springer Verlag. pp. 69–97.)

Appelt, Douglas E. (1985b). *Planning English Sentences.* Cambridge, UK: Cambridge University Press.

Austin, J. L. (1975). *How to Do Things With Words, Second Edition*. Cambridge, MA: Harvard University Press. (The first edition, which is much more widely cited, is copyright 1962).

Becker, Joseph D. (1975). "The Phrasal Lexicon." In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1.* Morristown, NJ: ACL. pp. 60–63.

Berwick, Robert C. (1983). "Introduction: Computational Aspects of Discourse." In Brady, Michael, and Berwick, Robert C., eds., (1983). *Computational Models of Discourse.* Cambridge, MA: MIT Press. pp. 27–105.

Block, Russell, and Horacek, Helmut (1990). "Generating Referring Expressions Using Multiple Knowledge Sources." *Proceedings of COLING-1990* (Helsinki, Finland). Morristown, NJ: ACL, pp. 24–29.

Brachman, Ronald (1979). "On the Epistemological Status of Semantic Networks." In Findler, Nicholas V., ed., *Associative Networks: Representation and Use of Knowledge by Computers.* New York: Academic Press. pp. 3–50. (Reprinted in Brachman, Ronald and Levesque, Hector, eds., 1985. *Readings in Knowledge Representation.* Los Altos, CA: Morgan Kaufmann. pp. 191–216.)

Brachman, Ronald, and Levesque, Hector, eds. (1985). *Readings in Knowledge Representation.* Los Altos, CA: Morgan Kaufmann.

Brady, Michael, and Berwick, Robert C., eds. (1983). *Computational Models of Discourse.* Cambridge, MA: MIT Press.

Brennan, S.E. (1991). "Conversation With and Through Computers." *User Modeling and User-Adapted Interaction.* 1, pp. 67–86.

Britton, Bruce K., and Black, John B. (1985). *Understanding Expository Text: A Theoretical and Practical Handbook for Analyzing Explanatory Text.* Hillsdale, NJ: Lawrence Erlbaum.

Callow, Kathleen, and Callow, John C. (1992). "Text as Purposive Communication: A Meaning-based Analysis." In Mann, William C., and Thompson, Sandra A., eds., 1992. *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text.* Philadelphia: John Benjamins. pp. 5–38).

Cawsey, Alison (1992). *Explanation and Interaction: The Computer Generation of Explanatory Dialogues.* Cambridge, MA: MIT Press. (ACL-MIT Press Series)

Chafe, Wallace L. (1975). "Some Thoughts on Schemata." In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1.* Morristown, NJ: ACL. pp. 89–91.

Charniak, Eugene (1975). "Organization and Inference in a Frame-like System of Common Sense Knowledge." In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1.* Morristown, NJ: ACL. pp. 42–51.

Chomsky, Noam (1988). *Language and Problems of Knowledge: The Managua Lectures.* Cambridge, MA: MIT Press.

Clark, Herbert H. (1975). “Bridging” In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1*. Morristown, NJ: ACL. pp. 169–174.

Clark, Herbert H., and Marshall, Catherine (1978). “Reference Diaries.” In Waltz, David L., ed., 1978, *TINLAP-2*. New York: ACM. pp. 57–63.

Clark, Herbert H. and Schaefer, E.F. (1989). “Contributing to Discourse.” *Cognitive Science*, 13, pp. 259–294.

Cohen, Philip R., Morgan, Jerry, and Pollack, Martha E., eds. (1990). *Intentions in Communication.* Cambridge, MA: MIT Press/Bradford Books.

Conklin, E. Jeffrey, and McDonald, David D. (1982). “Salience: The Key to the Selection Problem in Natural Language Generation.” *Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics*, Morristown, NJ: ACL, pp. 129–135.

Dreyfus, Hubert L. (1981). “From Micro-Worlds to Knowledge Representation: AI at an Impasse.” In Haugeland, John. ed., *Mind Design*, Cambridge, MA: MIT Press, pp. 161–204. (Reprinted in Brachman, Ronald, and Levesque, Hector, eds., 1985. *Readings in Knowledge Representation.* Los Altos, CA: Morgan Kaufmann. pp. 71–94.)

Dreyfus, Hubert L., and Dreyfus, Stuart E. (1988). “Making a Mind vs. Modeling the Brain: Artificial Intelligence at a Branchpoint.” In Graubard, Stephen R., ed., 1988, *The Artificial Intelligence Debate: False Starts and Real Foundations.* Cambridge, MA: MIT Press, pp. 15–43. That book is a reprint of: *Artificial Intelligence* (A Special Issue of) *Daedalus* 117(1).

Elhadad, Michael, and McKeown, Kathleen R. (1990). “Generating Connectives.” *Proceedings of COLING-1990* (Helsinki, Finland). Morristown, NJ: ACL, pp. 97–101.

Evens, Martha W. (1988). *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks.* Cambridge, UK: Cambridge University Press.

Evens, Martha W., Spitkovsky, John, Boyle, Patrick, Michael, Joel A., and Rovick, Allen A. (1993). “Synthesizing Tutorial Dialogues.” *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (Boulder, CO) Hillsdale, NJ: Erlbaum, pp. 137–142.

Fass, Warren, and Schumacher, Gary M. (1981). "Schema Theory and Prose Retention: Boundary Conditions for Encoding and Retrieval Effects." *Discourse Processes* 4, pp. 17–26.

Feldman, Jerry (1975). "Bad-Mouthing Frames." In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1*. Morristown, NJ: ACL. pp. 92–93.

Fetzer, James H. (1991). *Philosophy and Cognitive Science.* New York: Paragon House.

Fox, Barbara (1990). *Human Tutorial Dialogue: Final Report.* ONR Cognitive Science Program, Contractor/Grantee Meeting.

Fox, Barbara (1993). *The Human Tutorial Dialogue Project.* Hillsdale, NJ: Erlbaum.

Frohlich, David and Luff, Paul (1990). "Applying the Technology of Conversation to the Technology for Conversation." In Frohlich, David, Luff, Paul, and Gilbert, N., *Computers and Conversation.* London, UK: Academic Press, pp. 187–220.

Galdes, K.K., Smith, P.J., and Smith, J.W., Jr. (1991). "Factors Determining When to Interrupt and What to Say: An Empirical Study of the Case-Method Tutoring Approach." *Proceedings of the International Conference on the Learning Sciences,* (Evanston, IL), pp. 194–202.

Garfield, Jay L., ed. (1990). *Foundations of Cognitive Science: The Essential Readings.* New York: Paragon House.

Genesereth, Michael R., and Nilsson, Nils J. (1987). *Logical Foundations of Artificial Intelligence.* Los Altos, CA: Morgan Kaufmann.

Graesser, Arthur C., Lang, K., and Horgan, D. (1988). "A Taxonomy for Question Generation." *Questioning Exchange*, 2, pp. 3–15.

Graesser, Arthur C., and Franklin, Stanley P. (1990). "QUEST: A Cognitive Model of Question Answering." *Discourse Processes,* 13, pp. 279–303.

Graesser, Arthur C., and Hemphill, Darold (1991). "Question Answering in the Context of Scientific Mechanisms." *Journal of Memory and Language,* 30, pp. 186–209.

Graubard, Stephen R., ed. (1988). *Artificial Intelligence.* (a special issue of) *Daedalus* 117(1). Republished as *The Artificial Intelligence Debate: False Starts, Real Foundations.* Cambridge, MA: MIT Press, 1988.

Green, Nancy, and Carberry, Sandra (1994). "A Hybrid Model for Indirect Answers." *Proceedings of the Annual Meeting of the Association for Computational Linguistics.* (Las Cruces, NM). Morristown, NJ: Association for Computational Linguistics. pp. 58–65.

Grice, H. Paul (1975). "Logic and Conversation" In Cole, P. and Morgan, J.L., eds., *Speech Acts, Syntax and Semantics 3*, New York: Academic Press, pp. 41–58.

Grosz, Barbara J. (1977). "The Representation and Use of Focus in a System for Understanding Dialogs." *Proceedings of the Fifth International Joint Conference on Artificial Intelligence* (Cambridge MA, 1977). Los Altos, CA: Morgan Kaufmann, pp. 67–76.

Grosz, Barbara J. (1978). "Focusing In Dialog." In Waltz, David L., ed., 1978, *TINLAP-2.* New York: ACM. pp. 96–103.

Grosz, Barbara J., and Sidner, Candace L. (1986). "Attention, Intentions, and the Structure of Discourse." *Computational Linguistics*, 12(3), pp. 175–204.

Grosz, Barbara J., Appelt, Douglas, Martin, Paul, and Pereira, Fernando (1987). "TEAM: An Experiment in the Design of Transportable Natural Language Interfaces." *Artificial Intelligence*, 37(2), pp. 173–244.

Grosz, Barbara J., Pollack, Martha E., and Sidner, Candace L. (1989). "Discourse." In Posner, Michael, ed., *Foundations of Cognitive Science.* Cambridge, MA: MIT Press, 1989. pp. 437–468.

Hajicova, Eva, Sgall, Petr, and Kubon, Vladislav (1990). "Hierarchy of Salience and Discourse Analysis and Production." *Proceedings of COLING-1990* (Helsinki, Finland). Morristown, NJ: ACL, pp. 144–148.

Hayes, Patrick J. (1979). "The Logic of Frames." In Metzing, D., ed., *Frame Conceptions and Text Understanding.* Berlin: Walter de Gruyter and Co., pp. 46–61. (Reprinted in Brachman, Ronald, and Levesque, Hector, eds., 1985. *Readings in Knowledge Representation.* Los Altos, CA: Morgan Kaufmann. pp. 287–296.)

Hearst, Marti A. (1994). "Multi-paragraph Segmentation of Expository Text." *Proceedings of the Annual Meeting of the Association for Computational Linguistics.* (Las Cruces, NM). Morristown, NJ: Association for Computational Linguistics. pp. 9–16.

Hewitt, Carl (1975). "Stereotypes as an ACTOR Approach Towards Solving the Problem of Procedural Attachment in FRAME Theories." In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1.* Morristown, NJ: ACL. pp. 94–103.

Hirst, Graeme (1987). *Semantic Interpretation and the Resolution of Ambiguity.* Cambridge, UK: Cambridge University Press.

Hovy, Eduard H. (1988). "Planning Coherent Multisentential Text." *Proceedings of the 26th Annual Meeting of the ACL,* Morristown, NJ: ACL, pp. 163–169.

Hume, Gregory D., Michael, Joel A., Rovick, Allen A., and Evens, Martha W. (1993). "The Use of Hints as a Tutorial Tactic." *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum Assoc. pp. 563–568.

Hume, Gregory, Michael, Joel, Rovick, Allen, and Evens, Martha (1995). "Hinting as a Tactic in One-On-One Tutoring." *Journal of the Learning Sciences* (forthcoming)

Jackendoff, Ray (1975). "A System of Semantic Primitives." In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1.* Morristown, NJ: ACL. pp. 24–29.

Jacobs, Paul S. (1988). "PHRED: A Generator for Natural Language Interfaces." In McDonald, David D., and Bolc, Leonard, eds., 1988. *Natural Language Generation Systems.* New York: Springer Verlag. pp. 312–352.

Jordan, Michael P. (1992). "An Integrated Three-Pronged Analysis of a Fund-Raising Letter." In Mann, William C., and Thompson, Sandra A., eds., 1992. *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text.* Philadelphia: John Benjamins. pp. 171–226.

Kaufmann, David R., Patel, Vimla L., and Magder, Sheldon A. (1992). "Conceptual Understanding of Circulatory Physiology." *Annual Meeting of the American Education Research Association* (San Francisco, CA).

Kukich, Karen (1988). “Fluency in Natural Language Reports.” In McDonald, David D., and Bolc, Leonard, eds., 1988. *Natural Language Generation Systems.* New York: Springer Verlag. pp. 280–311.

Lakoff, George (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind.* Chicago: University of Chicago Press.

Lakoff, George, and Johnson, Mark (1980). *Metaphors We Live By.* Chicago: University of Chicago Press.

Lambert, Lynn, and Carberry, Sandra (1992). “Modeling Negotiation Subdialogues.” *30th Annual Meeting of the Association for Computational Linguistics,* Morristown, NJ: ACL, pp. 193–200.

Langacker, Ronald W. (1991). “Cognitive Grammar.” In Droste, F.G. and Joseph J.E., eds. (1991), *Linguistic Theory and Grammatical Description.* Amsterdam/Philadelphia: John Benjamins, pp. 275–306.

Levesque, Hector J. (1986). “Knowledge Representation and Reasoning.” In *Annual Review of Computer Science, Vol 1.*, pp. 255–287.

Levesque, Hector, and Brachman, Ronald (1985). “A Fundamental Tradeoff in Knowledge Representation and Reasoning (Revised Version).” Brachman, Ronald, and Levesque, Hector, eds., 1985. *Readings in Knowledge Representation.* Los Altos, CA: Morgan Kaufmann. pp. 41–70.

Li, Jun, Seu, Jai, Evens, Martha W., Michael, Joel A., and Rovick, Allen A. (1992). “Computer Dialogue System (CDS): A System for Capturing Computer-Mediated Dialogue.” *Behavior Research Methods, Instruments, and Computers.* 24, pp. 535–540.

Longacre, Robert E. (1992). “The Discourse Strategy of an Appeals Letter.” In Mann, William C., and Thompson, Sandra A., eds., 1992. *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text.* Philadelphia: John Benjamins. pp. 109–130.

Luperfoy, Susann (1992). “The Representation of Multimodal User Interface Dialogues Using Discourse Pegs.” *30th Annual Meeting of the Association for Computational Linguistics,* Morristown, NJ: ACL, pp. 22–31.

Maida, Anthony, and Shapiro, Stuart C. (1982). "Intensional Concepts in Propositional Semantic Networks." *Cognitive Science*. 6(4), pp. 291–330. (Reprinted in Brachman, Ronald, and Levesque, Hector, eds., 1985. *Readings in Knowledge Representation.* Los Altos, CA: Morgan Kaufmann. pp. 170–189.)

Mann, William C., and Thompson, Sandra A. (1986). "Relational Propositions in Discourse." *Discourse Processes*, 9, pp. 57–90.

Mann, William C., and Thompson, Sandra A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization.* Technical Report ISI/RS-87-190. Marina del Rey: University of Southern California/Information Sciences Institute. Reprinted from Polanyi, Livia, ed.(1987). *The Structure of Discourse*. Norwood, NJ: Ablex.

Mann, William C., and Thompson, Sandra A., eds. (1992). *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text.* Philadelphia: John Benjamins.

Mann, William C., Matthiessen, Christian M.I.M., and Thompson, Sandra A. (1992). "Rhetorical Structure Theory and Text Analysis." In Mann, and Thompson, eds., 1992. *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text.* Philadelphia: John Benjamins. pp. 39–78.

Marcus, Mitchell (1975). "Diagnosis as a Notion of Grammar." In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1.* Morristown, NJ: ACL. pp. 6–10.

Marcus, Mitchell (1980). *A Theory of Syntactic Recognition for Natural Language.* Cambridge, MA: MIT Press.

McCarthy, John (1977). "Epistemological Problems in Artificial Intelligence." *Proceedings of IJCAI-1977*, pp. 1038–1044. Los Altos, CA: Morgan Kaufmann. (Reprinted in Brachman, Ronald, and Levesque, Hector, eds., 1985. *Readings in Knowledge Representation.* Los Altos, CA: Morgan Kaufmann. pp. 23–30.)

McCarthy, John, and Hayes, Patrick J. (1969). "Some Philosophical Problems From the Standpoint of Artificial Intelligence." In Meltzer, B., and Michie, D., eds. (1969). *Machine Intelligence 4.* Edinburgh: Edinburgh University Press.

McCawley, James D. (1988). "Discourse Syntax." *The Syntactic Phenomena of English, Volume 2.* Chicago: University of Chicago Press, pp. 754–768.

McCoy, Kathleen F. (1989). "Generating Context-Sensitive Responses to Object-Related Misconceptions." *Artificial Intelligence*, 41, pp. 159–195.

McDonald, David D. (1983). "Natural Language Generation as a Computational Problem: an Introduction" In Brady, Michael, and Berwick, Robert C., eds., 1983. *Computational Models of Discourse.* Cambridge, MA: MIT Press. pp. 209–265.

McDonald, David D., and Bolc, Leonard, eds. (1988). *Natural Language Generation Systems.* New York: Springer Verlag.

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

Meyer, Bonnie J.F. (1985). "Prose Analysis: Purposes, Procedures, and Problems." In Britton and Black (1985). *Understanding Expository Text: A Theoretical and Practical Handbook for Analyzing Explanatory Text.* Hillsdale, NJ: Lawrence Erlbaum. pp. 11–64 and 269–304.

Meyer, Bonnie J.F. (1992). "An Analysis of a Plea for Money." In Mann, William C., and Thompson, Sandra A., eds., 1992. *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text.* Philadelphia: John Benjamins. pp. 79–108.

Minsky, Marvin (1975?). "Minsky's Frame System Theory." In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1.* Morristown, NJ: ACL. pp. 104–116.

Minsky, Marvin (1981). "A Framework for Representing Knowledge." In Haugeland, John, ed., *Mind Design*. Cambridge MA: MIT Press, pp. 95–128. (Reprinted in Brachman, Ronald, and Levesque, Hector, eds., 1985. *Readings in Knowledge Representation.* Los Altos, CA: Morgan Kaufmann. pp. 245–262.)

Minsky, Marvin (1985, 1986). *The Society of Mind.* New York: Simon and Schuster.

Moore, Johanna D. (1995). *Participating in Explanatory Dialogues.* Cambridge, MA: MIT Press. (ACL-MIT Press Series)

Moore, Johanna D., and Paris, Cecile L. (1989). "Planning Text for Advisory Dialogues." *Proceeding of the 27th Annual Meeting of the ACL*, Morristown, NJ: ACL, pp. 203–211.

Moore, Johanna D., and Paris, Cecile L. (December 1992). *Planning Text for Advisory Dialogues: Capturing Intentional and Rhetorical Structure.* Technical Report 92-22, University of Pittsburgh.

Moore, Johanna D., and Pollack, Martha E. (1992). "A Problem for RST: The Need for Multi-Level Discourse Analysis." *Computational Linguistics,* 18(4), December, 1992, pp. 537–544.

Nirenburg, Sergei, ed. (1987). *Machine Translation: Theoretical and Methodological Issues.* Cambridge, UK: Cambridge University Press.

Osherson, Daniel N., and Lasnik, Howard, eds. (1990). *Language: An Introduction to Cognitive Science, Vol. 1.* Cambridge, MA: MIT Press.

Partridge, Derek, and Wilks, Yorick (1990). *The Foundations of Artificial Intelligence: A Sourcebook.* Cambridge, UK: Cambridge University Press.

Pike, Evelyn G. (1992). "How I Understand a Text—via the Structure of the Happenings and the Telling of Them." In Mann, William C., and Thompson, Sandra A., eds., 1992. *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text.* Philadelphia: John Benjamins. pp. 227–261.

Pinker, Steven (1989). *Learnability and Cognition: The Acquisition of Argument Structure.* Cambridge, MA: MIT Press.

Pinker, Steven, and Prince, Alan (1988). "On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition." *Cognition* (28). (Reprinted in Pinker, Steven and Mehler, Jacques, eds. (1988), *Connections and Symbols*, Cambridge, MA: MIT Press, pp. 75–193.)

Pollack, Martha E. (1986). "A Model of Plan Inference that Distinguishes Between the Beliefs of Actors and Observers." *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics.* Morristown, NJ: Association for Computational Linguistics, pp. 207–214.

Pollard, Carl, and Sag, Ivan A. (1987). *Information-Based Syntax and Semantics: Volume 1, Fundamentals.* (CSLI Lecture Notes; Number 13). Stanford/Chicago: CSLI/University of Chicago Press.

Posner, Michael I., ed. (1989). *Foundations of Cognitive Science.* Cambridge, MA: MIT Press.

Putnam, Hilary (1988). *Representation and Reality.* Cambridge, MA: MIT Press.

Reichman, Rachel (1985). *Getting Computers to Talk Like You and Me: Discourse Context, Focus, and Semantics (An ATN Model).* Cambridge, MA: MIT Press.

Reiter, Raymond (1975). "Formal Reasoning and Language Understanding Systems." In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1.* Morristown, NJ: ACL. pp. 175–179.

Rieger, Chuck (1975). "The Commonsense Algorithm as a Basis for Computer Models of Human Memory, Inference, Belief and Contextual Language Comprehension." In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1.* Morristown, NJ: ACL. pp. 180–195.

Riesbeck, Christopher K. (1975). "Computational Understanding." In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1.* Morristown, NJ: ACL. pp. 11–16.

Rovick, Allen A. and Michael, Joel A (1992). "The Prediction Table: A Tool for Assessing Students' Knowledge." *(Advances in Physiology Education, 8) American Journal of Physiology,* 263, pp. S33–S36.

Salton, Gerard, and McGill, Michael J. (1983). *Introduction to Modern Information Retrieval.* New York: Computer Science Press/McGraw Hill.

Sanders, Gregory A., Evens, Martha W., Hume, Gregory D., Rovick, Allen A., and Michael, Joel A. (1992). "An Analysis of How Students Take the Initiative in Keyboard-to-keyboard Tutorial Dialogues in a Fixed Domain." *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society* (Bloomington, IN), Hillsdale, NJ: Lawrence A. Erlbaum Assoc., pp. 1086–1091.

Schank, Roger, and Nash-Webber, Bonnie L., eds. (1975). *Theoretical Issues in Natural Language Processing. An Interdisciplinary Workshop in Computational Linguistics, Psychology, and Artificial Intelligence.* (TINLAP-1). Morristown, NJ: ACL.

Searle, John R. (1969). *Speech Acts: An Essay in the Philosophy of Language.* Cambridge, UK: Cambridge University Press.

Searle, John R. (1975). "Indirect Speech Acts." In Cole, Peter, and Morgan, Jerry L., eds., *Syntax and Semantics, Volume 3: Speech Acts.* New York, NY: Academic Press.

Searle, John (1992). *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Sells, Peter (1985). *Lectures on Contemporary Syntactic Theories: An Introduction to Government-Binding Theory, Generalized Phrase Structure Grammar, and Lexical-Functional Grammar.* (CSLI Lecture Notes; Number 3). Stanford/Chicago: CSLI/University of Chicago Press.

Seu, Jai, Chang, Ru-Charn, Li, Jun, Evens, Martha W., Michael, Joel A., and Rovick, Allen A. (1991). "Language Differences in Face-to-Face and Keyboard-to-Keyboard Sessions." *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (Chicago, IL) Hillsdale, NJ: Erlbaum 576–580.

Shoham, Yoav (1987). "Temporal Logics in AI: Semantical and Ontological Considerations." *Artificial Intelligence*, 33(1), pp. 89–104.

Sidner, Candace Lee (1983). "Focusing in the Comprehension of Definite Anaphora." In Brady, Michael, and Berwick, Robert C., eds., 1983. *Computational Models of Discourse.* Cambridge, MA: MIT Press. pp. 267–330.

Smith, Brian C. (1982). Prologue to *Reflection and Semantics in a Procedural Language.* (This is the Prologue to Smith's PhD Dissertation) Reprinted in Brachman, Ronald, and Levesque, Hector, eds., 1985. *Readings in Knowledge Representation.* Los Altos, CA: Morgan Kaufmann. pp. 31–40.

Smith, Ronnie W., Hipp, D. Richard, and Biermann, Alan W. (1992). "A Dialog Control Algorithm and Its Performance." *Third Conference on Applied Natural Language Processing,* Morristown, NJ: ACL, pp. 9–16.

Sowa, John F. (1988). "Semantic Networks." In Shapiro, Stuart, ed., *Encyclopedia of Artificial Intelligence*, New York: John Wiley & Sons, pp. 1011–1024.

Stillings, Neil A., Feinstein, Mark H., Garfield, Jay L., Rissland, Edwina L., Rosenbaum, David A., Weisler, Steven E., and Baker-Ward, Lynne (1987). *Cognitive Science: An Introduction*. Cambridge, MA: MIT Press.

Traum, David R. and Allen, James F. (1994). "Discourse Obligations in Dialogue Processing." *Proceedings of the Annual Meeting of the Association for Computational Linguistics.* (Las Cruces, NM). Morristown, NJ: Association for Computational Linguistics. pp. 1–8.

Waltz, David L., ed. (1978). *Theoretical Issues In Natural Language Processing–2* (Urbana-Champaign IL). New York: ACM. — generally referred to as "*TINLAP-2*"

Webber, Bonnie Lynn (1983). "So What Can We Talk About Now?" In Brady, Michael, and Berwick, Robert C., eds., 1983. *Computational Models of Discourse.* Cambridge, MA: MIT Press. pp. 331–371.

Wick, M.R. and Thompson, W.B. (1992). "Reconstructive Expert System Explanation." *Artificial Intelligence* 54. pp. 33–70.

Wilks, Yorick (1975). "Primitives and Words." In Schank, Roger, and Nash-Webber, Bonnie L., eds., 1975, *TINLAP-1.* Morristown, NJ: ACL. pp. 38–41.

Winfrey, LaPearl Logan, and Goldfried, Marvin R. (1986). "Information Processing and the Human Change Process." In Ingram, Rick E. *Information Processing Approaches to Clinical Psychology.* New York, NY: Academic Press. pp. 241–258.

Winograd, Terry (1978). "On Primitives, Prototypes, and Other Semantic Anomalies." In Waltz, David L., ed., 1978, *TINLAP-2.* New York: ACM. pp. 25–32.

Winograd, Terry and Flores, Fernando (1987). *Understanding Computers and Cognition.* Reading, MA: Addison Wesley.

Winter, Eugene (1992). "The Notion of Unspecific versus Specific as one Way of Analysing the Information of a Fund-Raising Letter." In Mann, William C., and Thompson, Sandra A., eds., 1992. *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text.* Philadelphia: John Benjamins. pp. 131–170.

Woods, William (1975). "What's in a Link." In Bobrow, Daniel G., and Collins, Allan, eds., *Representation and Understanding: Studies in Cognitive Science.* New York: Academic Press, pp. 35–82. (Reprinted in Brachman, Ronald, and Levesque, Hector, eds., 1985. *Readings in Knowledge Representation.* Los Altos, CA: Morgan Kaufmann. pp. 217–242.)

Woolf, Beverly Park (1984). *Context Dependent Planning in a Machine Tutor.* PhD dissertation. University of Massachusetts at Amherst. (COINS Technical Report 84–21).

Zhang, Yuemei (1991). *Knowledge-Based Discourse Generation For an Intelligent Tutoring System.* (PhD Dissertation) Chicago, IL: Illinois Institute of Technology.