SRI International

COMPUTATIONAL STRATEGIES FOR ANALYZING THE ORGANIZATION AND USE OF INFORMATION

Technical Note 253

July 1981

Donald E. Walker
Artificial Intelligence Center
SRI International

Projects 8794, 1944, and 676D32

Prepared for the
National Institute of Education Project on
Knowledge Synthesis and Interpretation Processes

To be published in Knowledge Structure and Use:
Perspectives on Synthesis and Interpretation,
edited by Spencer Ward and Linda Reed.
Washington, D.C.: National Institute of Education,
in cooperation with CEMREL, Inc., St. Louis, Missouri,
1981.

SRI International®

333 Ravenswood Ave. • Menlo Park, CA 94025
(415) 859-6200 • TWX: 910-373-2046 • Telex: 334 486

ABSTRACT

This chapter describes new developments in computer-based procedures that can improve our understanding of how people organize and use information. Relevant recent research in information science, computational linguistics, and artificial intelligence is reviewed. A program of research is presented that is producing systems that make it possible to study the organization and use of information and, at the same time, provide more effective support for people engaged in those activities. Finally, several current projects that are part of this longer-term program are discussed.

CONTENTS

A. INTRODUCTION[1]

Every time a person learns something new, solves a problem, or makes a decision, he engages in an act of synthesis and interpretation. The complexity of the process and the significance of the results vary, but a central element in the activity is the identification of some new item as information and its integration into a body of knowledge that the person has accumulated in the course of his experiences. The body of knowledge may constitute the core of a scientific discipline, and the addition of the information may result in a creative insight with a revolutionary impact on the field. Or the knowledge may be quite informal, reflecting little on specific education or training, and the added information may simply allow the person to complete what he was doing without any clearly discernible effect on what he can be said to know. Between these extremes there is, of course, a broad range of variation, and people differ significantly in their behavior, in what constitutes "information" for them. In spite of the importance of these acts of synthesis and interpretation and their products, we know relatively little about the processes involved. We do not even have good strategies for studying those processes, much less for helping the people who are engaged in them perform more effectively.

It would not be possible or appropriate for me to consider all aspects of learning, problem solving, and decision making as they relate to knowledge synthesis and interpretation processes. My point of departure is what is usually called "information storage and retrieval"; my primary purpose will be to show how new developments in computer-based procedures for working with concepts of information, knowledge, and language can improve our understanding of how people organize and

---

[1] The preparation of this chapter was supported in part by grants from the National Cancer Institute (No. 1 R01 CA26655) and the National Library of Medicine (No. 1 R01 LM03611) and by SRI International Internal Research and Development funds. I would like to thank Susan Crawford, Spencer Ward, Nicholas Belkin, Linda Smith, Barbara Grosz, Staffan Loef, Robert Amsler, and Norman Haas for their helpful comments on the manuscript; my colleagues at SRI and elsewhere for their inspiration and counsel; and especially Hans Karlgren and Martin Epstein for their cooperation and their participation in my intellectual life.

use information. I will be concerned particularly with how data elements or text passages that are relevant for a person's needs in a particular situation can be located in a computer database, recovered from it, and applied to solve a current problem.

For the purposes of this chapter, it will be helpful to think of the person engaged in knowledge synthesis and interpretation as a scientist or scholar.[2] This constraint will allow me to restrict my concerns to data or texts that in some sense constitute formal communications intended to be shared with others and to be accumulated as having continuing relevance for people working in a given field. The person's need may be clear and able to be stated precisely as a hypothesis or question; or it may be vague initially, and specifiable only after it is satisfied. In either case the need is usually communicated, at least to other people and often as the person himself thinks of it, in natural language, for example, as a statement in a particular language, like English, or in some jargon that is derivative from such a language. The stored materials considered in relation to the need are also frequently formulated in natural language, although there may be specialized symbol systems, like mathematics or graphic notations, and conventionalized arrangements in tabular or graphic form. And, finally, the new synthesis and interpretation may be expressed in language and stored in a database for others to consider.

The circularity or cumulative nature of this process of deriving information and synthesizing knowledge has led people to describe the databases themselves as bodies of information or knowledge. Although a database can be used in a similar way by people with similar needs and backgrounds, it is important to recognize that their problems vary and that the knowledge in a discipline is changing constantly. The differences in needs and backgrounds may be subtle, but they affect how a person interprets a given set of data. Because my primary interest is in the nature of these differences, and because I want to stress the

---

[2] The terms "scientist" and "scholar" should be interpreted broadly to include any class of professionals, like physicians, lawyers, educators, and journalists.

process of interpretation rather than its product, I will be using the term information to identify the contents of the data elements or text passages that are determined by a person to be relevant to his need. It becomes appropriate to talk about knowledge, where the needs are shared by a group of people and there is a consensus about how bodies of data and text relate to those needs. The implications of this choice of definitions should become clear by the end of this chapter. However, the reader should be warned that these terms are used in other ways in the literature, and that conventional usages may not always allow me to be consistent.

Our society is respectful of information and knowledge; we preserve the source materials in articles and books, place them in libraries and special databases, and with increasing frequency store them in computers. There are a remarkably large number of computer-readable databases available already, and a range of systems for accessing them (Williams 1979; Bourne 1980; Hall and Brown 1981). In addition, more and more books, newspapers, and journals are being photocomposed from computerized source files. Word processing techniques have transformed the processes of preparing letters and other types of business correspondence, although these communications are likely to be stored in computer form, if at all, only for archival purposes. Networks now link computers together and provide remote access through terminals to increasing numbers of users. The developments in the last ten years have been sufficiently dramatic that Lancaster (1978), in his book Towards Paperless Information Systems, argues that computer storage and network access will replace conventional printing and publishing practices--and soon. Whether or not one agrees with this prediction, and I do consider it reasonable, it is clear that access to this broad range of source materials through computers will have an increasing impact on the way that people synthesize and interpret knowledge in the future. However, the existing procedures for selectively retrieving items from these databases are grossly inefficient, and if we are to realize the potential value of computers, we need to develop new access methods.

In the next section I will review recent developments in three fields in which the concepts of information, language, and knowledge are central concerns: information science, computational linguistics, and artificial intelligence. Then I will describe a research program, deriving from and building on those developments, which is designed to lead to computer systems that will make it easier to study how people organize and use information and, at the same time, provide more effective support for those who are engaged in these activities. Finally, I will discuss several current projects that are part of this longer-term program.

B. RECENT DEVELOPMENTS IN INFORMATION SCIENCE, COMPUTATIONAL LINGUISTICS, AND ARTIFICIAL INTELLIGENCE

Information science, computational linguistics, and artificial intelligence are all relatively new areas of research, each having assumed an independent identity within the past 25 years. They owe their emergence as recognized disciplines in substantial measure to the development of computers. Information science grew out of library science as an attempt to formalize procedures for storing and retrieving documents and for providing access to the information they contain. Implementing these procedures on computers has provided both theoretical and practical guidance for work in the field. Computational linguistics developed out of a multi-disciplinary interest in language and in the formulation of computer programs that perform human language skills. Artificial intelligence emerged from computer science, itself a relatively young field, reflecting the intent to model computationally the knowledge and behavior people use in carrying out tasks.

Although information, language, and knowledge are, respectively, central concepts for these three fields, they are not the exclusive property of any one of them. Moreover, there is an increasing amount of communication among people doing research in these fields and substantial involvement in exploring common areas of interest.[3]

Nevertheless, information science, computational linguistics, and artificial intelligence are recognized as independent for historical and sociological reasons, and in the discussion that follows it will be appropriate to consider each in turn as it relates to knowledge synthesis and interpretation.

Within information science, procedures have been developed for storing and retrieving data and documents, the primary source materials for knowledge synthesis and interpretation. The problems encountered and some of the directions along which solutions are being sought will be considered. Particularly relevant are some recent attempts to provide a definition of information for the field that reflects more adequately the critical role of the needs of users.

Computational linguistics is relevant for knowledge synthesis and interpretation in two ways: First, this discipline can provide practical techniques for interacting with and controlling the operations of a computer through natural language. By making access to computer-based data and text files conversational, we can both increase their utility and make it easier to observe how people work with them. Second, computational linguistics addresses the general issue of communication in natural language; recent research has made it clear how complex the processes associated with human understanding really are and how much more there is that we need to know. As our knowledge increases, the sophistication of our systems can be enhanced.

A key area of interest in artificial intelligence is knowledge representation, particularly as it is reflected in knowledge-based

---

[3] Cognitive science also intersects with the fields being considered in this chapter. In cognitive science, the primary reference point is psychology, and the focus is on formalizing and modeling human processes and capabilities; computer implementations constitute both tests and demonstrations of the results. Cognitive science is not being considered here because the developments I am reporting on are not directly motivated by psychological considerations. Nevertheless, some of the research I will attribute to artificial intelligence and computational linguistics would be identified by some people as belonging properly to cognitive science. See Collins (1977) for a position statement made for the inauguration of the journal Cognitive Science.

systems, which attempt to model the behavior of experts in a particular field. These systems actually contain knowledge syntheses at a high level of sophistication. Understanding how to represent knowledge adequately and how to make inferences based on it are essential for the development of capabilities that can support people who are engaged in synthesis and interpretation.

1. Information Science

Individual acts of synthesis and interpretation usually take place in a delimited context within a particular field of inquiry. In contrast, information science is concerned with identifying and formalizing that part of the process of knowledge formulation, organization, codification, retrieval, dissemination, and acquisition that is common to different disciplines. While it has been characterized as "having to do with storage, retrieval, and transmission of information of any kind, in any way" (Sparck Jones and Kay 1973, p. 2), it is more appropriate to limit our scope to records and documents, data and text files.[4] While the succession of Annual Reviews of Information Science and Technology (e.g., Williams 1980) illustrates the range of topics covered in the field, it is more helpful to relate it to a context established by what has been called the "information transfer cycle" (Lancaster 1979).

People prepare documents; they are published or otherwise printed in some less formal way; distribution is made either directly to potential users or indirectly to them through libraries and information centers; the documents are read by the user community, which includes the creators themselves; and their contents form one of the bases for

---

[4] Informal communications among specialists, particularly of the kind referred to as taking place within "invisible colleges" (Crane 1972), are obviously also of interest. Where these communications take the form of messages that are transcribed or otherwise captured in documentary form, they can be considered as objects for storage and retrieval. However, the special problems associated with spoken communications make it necessary to defer considering them until procedures for storing and searching acoustical recordings are better developed.

preparing new documents. With respect to this cycle, information-retrieval systems must deal necessarily with the acquisition and storage of materials, with procedures for organizing and controlling them, and with their delivery to particular users. The focus here will be on organization and control, and particularly on cataloging and classification on the one hand and requesting and searching on the other.

The major emphasis in this area has been and continues to be on bibliographic material, helping the user to identify primary or source documents that might have information relevant to his needs and interests (see Vickery 1961; Kochen 1974; Lancaster 1979; Van Rijsbergen 1979). As already noted, the number of bibliographic databases available online for this purpose is increasing steadily (Williams 1979; Hall and Brown 1981). However, current systems for accessing them, like Lockheed's DIALOG, the System Development Corporation's ORBIT, and MEDLARS and MEDLINE developed by the National Library of Medicine, still only provide pointers to the literature. Although the user can make a preliminary assessment of the utility of a document from its title and an abstract, the actual text itself must be located and evaluated to determine whether the contents really are relevant.

In contrast, there are systems that enable the user to access source data directly: either the complete texts of documents or a variety of different kinds of tabular values, both numerical and textual. For example, LEXIS (Mead Data Central) and WESTLAW (West Publishing Company) store millions of characters of legal statutes and court decisions that are available for searching (Sprowl 1976; Larson and Williams 1980). Economic data collected annually on countries throughout the world are provided through BISYSTEM (Business International Corporation). CRIB, a Computerized Resources Information Bank supported by the U.S. Geological Survey, contains data on mineral resources in a variety of countries. The information science community is becoming increasingly interested in source-data systems (cf. Landau et al. 1979; Wanger and Landau 1980).

Systems providing indirect reference to bibliographic data and systems providing direct access to source data have both become major resources for information specialists. However, there are problems encountered in conducting more complex searches that become especially critical when these systems are used for the kinds of interactions entailed in knowledge synthesis and interpretation. It will be useful to consider three areas in which difficulties arise. The first relates to the procedures for representing the content of a document, the second to the formulation of a request, and the third to the methods for relating the request to the documents in the collection.

a. Representing Document Content

Assigning a representation to a document for inclusion in a bibliographic database requires that a judgment be made about its content. Traditionally, as in a hierarchical classification like the Library of Congress or Dewey Decimal systems, a librarian or documentalist assigns a single code. While this procedure may be adequate as a general access point, particularly for a book, if a large number of items is assigned to the same category, a single identifier does not provide any basis for discriminating among them. Extending the hierarchies to improve discrimination provides only temporary help, since new developments require further extensions. More importantly, this approach is not appropriate where a document considers a variety of topics, which is the typical case.

To handle multiple subject relevance, most retrieval systems assign several index terms or thesaurus entries to a document, usually based on comprehensive term lists or a thesaurus established for a particular subject area. The assignments are most often made manually and by subject matter specialists, frequently on the basis of the title and abstract alone. Recent attempts to index documents automatically by selecting words from the texts themselves are providing promising results, but no operational systems use such techniques (Sparck Jones and Bates 1977; Harter 1978). As an alternative, Cooper and Maron

(1978) advocate a probabilistic and utility-theoretic approach to indexing, based on judgments about the likelihood that a document will satisfy a user. While potentially more discriminating, such a technique would require extremely sophisticated judgments. However the indexing is accomplished--and its effectiveness depends critically on the skill of the indexer or the indexing program, current techniques are conservative and do not provide for subtle discriminations. In addition, although new developments in a subject may require adding new entries to the term list, it is difficult if not impossible to reclassify the older documents to show these changes. Of course, if automatic indexing proves to be feasible, it could be used to process the entire file and update assignments as a field changes.

These problems associated with content representation for bibliographic retrieval are not encountered in source-data retrieval. Classification and indexing may not be necessary for full-text retrieval, but novel needs, changing terminology, and the possibility of paraphrasing a given topic in many different ways does complicate the identification of a potentially relevant passage. Working with bodies of tabular data is simpler, but it may be essential for the user to know a lot about the organization of the data and even the labels assigned to the subfiles and attributes. In any case, the problems encountered in request formulation and searching prove to be relevant for both kinds of retrieval systems (see O'Connor 1980), as the following discussion will show.

b. Formulating a Request

The formulation of the request is the second area in which difficulties are encountered in information storage and retrieval. It is important to ensure that most of the relevant material in the file is retrieved ("high recall") and that most of the material retrieved is relevant ("high precision"). Providing a powerful query language is a critical requirement for this purpose, particularly when the document collection is large and relatively homogeneous. Most query languages

provide a framework within which a set of identifiers and terms that identify the area of interest can be specified. This set is then matched against the corresponding representation assigned to the document in the database. The user selects an attribute, like author, and specifies a value for it, like "Newell" or "Simon". The use of Boolean logic makes it possible to constrain the relations of terms in various ways. Three Boolean operators are used most frequently: conjunction, disjunction, and negation. Conjunction is equivalent to the use of the term "and": "computers AND language" retrieves documents containing both those terms. Disjunction is equivalent to "or"; "computers OR language" retrieves documents containing either term. Negation is equivalent to "not"; it allows rejecting a document if it contains a specified term. Combining and grouping these operators makes it possible to request documents that contain "(computers AND language) OR (computational linguistics AND NOT artificial intelligence)"; that is, the documents retrieved would either have both "computers" and "language" or they would have "computational linguistics" but if they also had "artificial intelligence" they would not be retrieved. The effective use of query languages also requires the ability to specify synonyms and other kinds of related terms and to introduce truncated forms (like "comput", which will match "computer," "computers," "computing," and "computation") so that the search is not necessarily limited to one particular form of a word. A variety of aids are provided by different systems to help the user formulate a request (cf. Martin 1974; Lancaster 1979; Van Rijsbergen 1979).

The query languages used in systems for full-text searching are similar in many respects to those for bibliographic searching (Lancaster and Fayen 1973; Sprowl 1976). Thus, in legal searching one can specify the value for a particular field, like the title or citation for a case, the cases decided by a particular judge, or a state statute in a specified subject area, with truncation used to compensate for differences in the grammatical form of a word. Boolean searches can be made on words that occur in the text. However, it is also possible to constrain the search to sets of words in the same

passage or paragraph, rather than just anywhere in the document, and even to specify that they must occur within so many words of each other. Using the full text of a document eliminates (or at least reduces) the requirement for classification or indexing by a librarian or documentalist, but the user still has to formulate his request so that it can match the large variety of ways in which a particular idea can be expressed.

In both bibliographic and source-data retrieval, as the query language is made more complicated to increase power and discrimination, it becomes more difficult to use and certainly to remember, particularly for those who interact with a system only occasionally. As a result, a class of specialists has emerged to establish the requirements of the users and adapt them to the characteristics of the systems and of the databases they contain. These "information intermediaries" have proved to be extremely useful, particularly where the person with the request, the so-called "end user," can specify clearly what information is needed. The intermediary, working with the end user, helps in the formulation of a request by explaining what materials are available and how they are organized. However, the specialist, no matter how well qualified, cannot provide the depth of discrimination of the end user. Without such guidance, and in cases where the request can only be specified vaguely or in general terms, employing an intermediary inevitably results in retrieving an excessive amount of material, much of it unrelated to the original need.

A variety of strategies have been introduced to make request formulation easier; two are of particular interest here. For document or data collections that are arranged hierarchically, techniques have been developed to lead the user systematically along various pathways through the collection by presenting alternatives in the form of a menu from which successive choices are made (Fox and Palay 1979; Robertson, McCracken, and Newell 1979). For example, if the database contains a set of documents on information science,

computational linguistics, and artificial intelligence, the user would first be asked to choose one. Under information science, the selection might be planning information systems and services, basic techniques and tools, applications, the profession, and special topics.[5] Further alternatives are provided under each successive class, down to the level of the documents themselves, and, of course, a specific document may be included in several different classes. Each set of choices in the menu is presented as a frame; the network of frames establishes all of the questions that can be asked and all of the data that can answer them. Such a system requires a substantial amount of analysis to establish the frames, and, consequently, may not be appropriate for a database that changes frequently. One observation frequently made about menu-based systems is that although well suited to the novice who needs guidance, they can be frustrating to the experienced user. To compensate, some systems have introduced procedures that allow a person to proceed directly to a given area if the name is known, resorting to subsequent frames for refinement.

A second strategy for making information-retrieval systems easier to use is to allow a person to express his request in ordinary conversational language. This approach has been used primarily in what have been called "question-answering systems" for retrieving data from formatted files (Minker 1977). It makes use of a set of computational linguistic techniques which will be described in some detail in the section on Natural-Language Interface Systems below. For the present purposes it will be sufficient to remark that the request is analyzed in relation to a set of grammatical rules so that its syntactic and semantic structures are identified and the results translated into operations on the data. Natural-language queries can actually define a more complicated request than can be accomplished through Boolean operations on search terms. For example, it can distinguish readily between "What computer languages are used for translating programs?" and "What computer programs are used for translating languages?"

---

[5] Headings used to group chapters in the Annual Review of Information Science and Technology (see Williams 1980).

However, as will become clearer in the next section, the grammars required can become relatively complex. In addition, it is not always easy to describe to the user what linguistic constructions are allowable, since none of the grammars yet written covers more than a subset of English, although, as indicated in the next section, the scope of some contemporary grammars is relatively large. Watt (1968) introduced the concept of "habitability" to characterize the problems encountered by the user in accommodating to these limitations. Considerable progress has been made since then in writing grammars for a particular task domain and some work has been done on clarification dialogues that help the user specify the query (Codd 1974, 1978), but much more is required to provide the proper ease of accessibility.

An application of natural-language querying to bibliographic retrieval has just been developed that makes use of a much simpler mechanism. The person describes his interests in a few sentences or a paragraph; the words in the paragraph that are also index terms for the collection are identified and automatically converted into a more traditional query language form which is then used to search the database in the usual manner (Doszkocs and Rapp 1979). How far this approach can be carried remains to be determined, but the use of natural-language capabilities in information science certainly is likely to increase. The more general issues underlying natural-language processing will be addressed in the section on computational linguistics.

c. Relating the Request to the Document Representations

In discussing difficulties associated with information-retrieval systems, representing the content of items in the database and formulating requests have been considered separately. While this manner of presentation is appropriate for an introductory perspective, it is essential to recognize that within a particular system these two aspects are interdependent. A query language must be designed so that it allows specifying elements that can be matched against the document

representations. Correspondingly, indexing assignments should be made so that they reflect the way that requests are constructed. However well this coordination is accomplished at the design level, there are still difficulties associated with the matching process. A number of different strategies that are currently being explored show considerable promise for improving system performance (McGill and Huitfeldt 1979): weighting terms in relation to their frequency of occurrence in documents about a particular topic; clustering documents and terms on the basis of their cooccurrence; providing search keys to improve access to sets of documents; and introducing feedback procedures that allow the user to revise his request.

The strategies for improving system performance just mentioned all involve computational solutions. It is interesting in this context to note some recent work by Bates (1979a,b) based on studies of the psychological processes people actually engage in during searching. She has identified 29 "information search tactics" that can be used to keep the search on the track and efficient, to thread through the file structure, to design and redesign the search formulation, and to aid in the selection of specific terms. In addition, Bates has proposed 17 "idea tactics," to help generate new ideas or approaches to problems encountered in searching and to break out of unproductive patterns. The results of her analyses show the complexity of the interactions that are required in order to use current information-retrieval systems, and certainly reflect on the state of the technology. However, they also raise more general questions about information searching that reflect on the the nature of information itself, a topic that must be considered in more detail.

d. The Concept of Information in Information Science

In spite of the undeniable utility of information-retrieval systems, the conceptual foundations of information science are not well established. In particular, there is no agreement about the nature of "information" itself. The differences in interpretation

depend in part on stressing different aspects of the process of developing, distributing, and using documents and data. Approached in relation to production, information can be identified with the source materials that an author/originator creates or establishes. Viewed from the standpoint of embodiment in a database, information can be identified with the content of those materials, particularly as they are considered to be part of the cumulative store of "knowledge" for people in a discipline. Finally, information can be identified with the user, reflecting on what he determines to be relevant for his own needs and purposes. Both the "information transfer cycle" referred to in the introduction to this section and the notion of "information transfer" itself contain the tacit assumption that these three perspectives on information refer to the same object which is conveyed, essentially unchanged, from one person to another. Phrased in a communication theoretic form, an author could be said to encode his insights and observations into a message which is then transmitted through a channel to a receiver who extracts the content of the message, adding it to his own store of knowledge.

On the theoretical side, this view is supported by the notion of science (and indeed of "civilization" more generally) as an accumulation of knowledge with exponential growth characteristics (Price 1961). In spite of the occurrence of scientific revolutions (pace Kuhn 1970), the progressive nature of science, the day-to-day activities of scientists, and the operation of the information transfer cycle itself all testify to the aptness of the communication paradigm. Problems associated with the process are attributed to errors in encoding (lack of effective communication skills), difficulties with the channel (inadequate procedures for storing, accessing, and notifying), and to inaccuracies in decoding (lack of effective comprehension skills).

From a practical side, the information transfer view is also supported by information-retrieval technology. It is possible both to store and to extract relevant material, and a substantial industry has been developed that provides such services. That the services are

less than adequate is acknowledged, but this condition is considered to reflect on the immaturity of the field and not necessarily on the adequacy of its foundations.

In spite of the apparent consensus, this view of information is untenable. The recent review by Belkin (1978) of different information concepts that have been proposed for information science is particularly useful in this context. He has provided a particularly thorough examination of the approaches that have been taken to establish a coherent foundation for the field. He identifies the core problem of information science to be "facilitating the effective communication of desired information between human generator and human user" (p. 58). From this base, he sets out certain requirements that must be satisfied by an information concept, further categorizing them as definitional, behavioural, and methodological, on the one hand, and related to relevance and operational status, on the other (p. 62):

(1) It must refer to information within the context of purposeful, meaningful communication.
(2) It should account for information as a social communication process among human beings.
(3) It should account for information's being requested or desired.
(4) It should account for the effect of information on the recipient.
(5) It must account for the relationship between information and state of knowledge (of generator and of recipient).
(6) It should account for the varying effects of messages presented in different ways.
(7) It must be generalizable beyond the individual case.
(8) It should offer a means for prediction of the effects of information.

Belkin finds that few of the concepts that have been proposed in information science even begin to satisfy these requirements. His own approach begins with the user's recognition of an anomalous state of knowledge (see Belkin 1980); that state prompts the user to search for documents or other sources that might resolve the anomaly; the data retrieved are interpreted in relation to the user's conceptual framework in order to establish their underlying structure; if that structure seems to resolve the anomaly, the process is complete;

if not, the process begins again (see Belkin et al. 1979 Oddy 1977). The critical point here is that the relevance of the document for the user does not have to bear any necessary relation to the relevance its contents had for the person who generated it.

Reconciling the perspectives of the generator and the user is critical for future research in information retrieval, because it is precisely where there is a difference between them that information retrieval strategies have been called into question. Procedures for information retrieval have been developed to respond to a broad range of applications. Those procedures that depend on thesauruses and fixed sets of index terms work well with narrowly constrained bodies of data and where the user is interested in finding material that relates to one of a set of canonical problems that are standard for a given field of specialization.

In contrast, the situation most characteristic of knowledge synthesis and interpretation relates primarily to a class of users who, as professionals in a field, have complex problems for which there are no ready-made answers in a database. Relevant information has to be derived from the materials available. Moreover, the results of these derivations could be useful additions to the database because of their value for subsequent users. However, the way subsequent users would interpret these new materials would still reflect their own needs and purposes. In any dynamic field of knowledge, the problems and the information relevant to their solution are changing. Ways of accommodating these changes are essential. It is precisely for this reason that developments in computational linguistics and artificial intelligence are relevant. Furthermore, I believe that the research program described later in this chapter can be helpful in providing more responsive information-retrieval systems. Hopefully, it also will constitute a basis for elaborating and testing a more complex concept of information in which the perspectives of the producer, the discipline in general, and the particular user can all be incorporated.

The problems of establishing an information concept and reconciling the tension between the originator and the user of a body of knowledge are not unique to information science. Similar concerns have been advanced in science (Bronowski 1969), philosophy (Palmer 1969; Gadamer 1976), communication (Derwin, this volume), social policy (Schoen 1979), art (Cohen 1979), linguistics (Reddy 1979), and computational linguistics (Winograd 1980). They all seem to be part of an appropriate concern with how to understand the factors involved in the emergence of new developments in various fields of inquiry.

2. Computational Linguistics

A substantial part of the cumulative store of knowledge is embodied in natural-language form. People also tend to express the hypotheses and questions they have about that knowledge in the same way. Accordingly, it is essential to discuss the research that has been directed toward analyzing, processing, and even "understanding" language. The field of linguistics, of course, has that objective as its primary focus. However, in the context of a concern with information storage and retrieval systems, it is more appropriate to consider work in computational linguistics instead. Computational linguistics addresses many of the same topics as linguistics, but specifically from the standpoint of algorithmic formalization and computer modeling. In addition, it is concerned with a range of applications of natural-language processing, some of which are particularly relevant for information storage and retrieval.[6] Two directions being taken in recent work are of particular interest here: the development of question-answering and natural-language interface systems, and research on understanding natural-language discourse.

In the discussion of information-retrieval systems in the previous section, one of the strategies for making the formulation of

---

[6] Damerau (1976) made the last general survey of the field. The American Journal of Computational Linguistics is the only periodically devoted exclusively to this subject, although papers also appear in publications on computer science, artificial intelligence, linguistics, cognitive science, psychology, and literary analysis.

requests easier was allowing them to be stated in natural language. In computational linguistics there has been a long history of work on the development of such capabilities, particularly in the context of systems for question answering and fact retrieval. Recently, however, this approach has been extended, and more general interface facilities developed so that a person can access, interact with, and control a variety of different kinds of computer databases, text files, and software, expressing requests and commands in natural language. While these applications of computational linguistics do provide practical tools for information retrieval, their limitations reveal the problems that need to be addressed to determine how language is actually used in human communication. Accordingly, the central research effort in the field is concerned with clarifying the factors involved in natural-language understanding, particularly in the context of discourse and dialogue constraints. The process of hypothesis formation and testing certainly constitutes an example of a dialogue interaction. Since computer implementations play a key role in testing concepts, there have been a number of systems developed that can actually participate in conversations. The sections that follow consider these two directions in turn.

a. Question-Answering and Natural-Language Interface Systems

The initial achievement in computational linguistics was the design and implementation of parsers, that is, procedures for analyzing the syntactic patterns of sentences. Most of this early work was motivated by linguistic considerations, in particular to test grammatical theories, but a major effort from the beginning entailed querying files and retrieving data. This practical orientation reflected in part the difficulties encountered in translating linguistic concepts into computational form, but an equally important factor was the necessity for providing some context within which those concepts could be tested. The systems first developed for question answering or "fact retrieval", as it was also called to distinguish it from bibliographic retrieval, consisted of a parser and a set of

interpretation rules that converted the output of the parser into operations on a database.[7]

Parsers are guided by sets of grammatical rules that define acceptable relations among grammatical elements for the class of sentences defined by those rules. Parsing a sentence provides a set of analyses or "structural descriptions" that indicate the ways that the words it contains can be grouped to form a syntactically valid interpretation. There may be more than one such interpretation, even for sentences that appear unambiguous, at least when presented in context. A classic illustration is provided by the sentence "I saw the man in the park with the telescope" which has a dozen or more analyses (depending on the actual set of grammatical rules), reflecting the fact that "saw" can refer to cutting as well as seeing, the possibility that the telescope can be associated with "I" or with "the man," that either could be in the park, and so on. Procedures for semantic analysis can resolve some of these ambiguities--acts of sawing do not take place with ordinary telescopes, for instance, but to resolve other ambiguities it is necessary to appeal to knowledge shared by participants in a conversation and to previous remarks they have made. Thus, if the park referred to does not have a resident telescope, that instrument would be identified as having been brought by the man. If either the person speaking (the "I") or "the man" was known to be in the park, that ambiguity in the syntax would not be recognized. These and related issues involve discourse constraints that basic research on language understanding is just beginning to identify; their consideration will be deferred until the next section. In the development of question-answering systems, the concern has been almost exclusively with the "literal meaning" of a sentence, what could be determined by the analysis of the particular sequence of words in relation to whatever database established the set of possible meanings. Providing literal meaning certainly has proved to be difficult enough.

---

[7] For representative reviews of this early work, see Simmons (1970) and Walker (1973).

In a typical question-answering system, once the parser has provided a structural description, the interpretation rules identify the logical connections among the linguistic elements that correspond to database entries, and the corresponding retrieval operations are performed. An example may make this clearer. In the LUNAR system (Woods et al. 1972), which contained chemical data on lunar rock and soil composition from the Apollo moon missions, a geologist could ask "What is the average concentration of aluminum in high alkali rocks?" The parser would establish a syntactic structure that identified this sentence as consisting of a question word followed by a form of the word "be" and then a predicate complement construction consisting of a noun phrase (containing an adjective "average" modifying a noun "concentration") followed by two prepositional phrases. The phrase "high alkali rocks" corresponded to a particular set of entries in the database, and "aluminum" identified one of the attributes. The interpretation rules, applied to this structure, would translate the noun phrase "average concentration" into a particular set of computations on the values for the specified element "aluminum" that occurs in that particular kind of rock, and the geologist would receive the appropriate answer.

LUNAR was one of the most sophisticated of the early question-answering systems, as well as one of the earliest to be tested by specialists in the subject matter it dealt with. However, its limitations were clearly recognized by its developers. Although its grammar was extensive, it still covered only a relatively small subset of English, and it was not easy to specify to a user just what constructions were allowed. It could handle some pronouns and definite determiners, establishing a limited dialogue capability; thus, after processing the example sentence above, it could also analyze "What is its average concentration in igneous rocks?", relating "it" to "aluminum." LUNAR's knowledge about the use of its grammatical constructions and about the meanings of words and phrases was constrained by the particular database and by the ways selected for processing the data. The inference rules incorporated were

correspondingly restricted. In spite of these limitations, LUNAR and the other early systems should be considered significant accomplishments. In addition, they established the complexity of the problems involved and the necessity for sophisticated procedures for syntactic analysis, semantics, and inference.

In spite of the fact that we do not yet have general answers to many of these problems, recent research has led to a number of systems that are being used to provide natural-language interfaces. They make it possible to interact easily and directly with a database and other software by means of English questions, commands, and statements. All of these systems capitalize on the predictable and limited range of queries required to access the associated files or perform the repertory of commands, and their success depends critically on the relatively small number of operations entailed in the interpretation phase. There are two major approaches being taken for these systems, distinguished primarily by the level of generality of the grammar and the degree to which the grammatical analysis and interpretation phases are integrated for a particular application.

The LIFER system (Hendrix 1977a,b; Hendrix et al. 1978) provides an example of a natural-language interface with a special-purpose grammar that closely couples analysis and interpretation.[8] In order to be able to build interfaces more rapidly, it makes use of a model of grammar that is more ad hoc and less linguistically motivated than the one described above. Rather than use categories like "noun phrase" and "verb phrase," semantic characteristics are incorporated directly into the grammatical rules (Brown and Burton 1975). Thus, a semantic grammar might contain "how-many," "ship-attribute," "disease-type", and "tabulate-clause." Rules formed from these categories simplify the process of translating the query into operations on the database, since restrictions on the meaningful relations among concepts are already built in.

---

[8] Systems along these lines have also been developed by Harris (1977, 1979), Templeton (1979), and Waltz (1978a); LIFER is chosen for illustration here because two of the projects in the research section at the end of this chapter make use of it.

LIFER furnishes the interface builder with a sophisticated set of tools. Language-specification routines simplify the process of developing rules by making it easy to: (a) create and test grammatical patterns; (2) establish word classes and associate words with them; (3) group together words in "fixed phrases" that should be analyzed as a whole; and (4) introduce the predicates needed to perform operations on the associated database or in the software to which it is linked. LIFER also includes a spelling corrector, an automatic facility for processing incomplete ("elliptical") sentences by relating them to requests that have already been analyzed, and a paraphrase mechanism that allows the user to extend the language interface by defining new constructs at the word, phrase, and sentence level.

LIFER makes it possible to build a simple language interface for a new application easily and to develop relatively powerful and efficient systems in proportion to the effort expended. If a reasonably complete subset of syntactic constructions is required for comfortable use in a given application, extending the grammar can consume several years of effort (Hendrix et al. 1978). Such an expenditure may be justifiable in a given case, but since the resulting system is dependent on the subject matter of the particular problem domain selected, a new application will require writing a new grammar and a new set of interpretation rules.

In contrast to the domain-dependence of LIFER, the DIAMOND/DIAGRAM system (Robinson 1980; Robinson et al. 1980) provides a language interface capability designed for linguistic generality and ease of applicability to new problem areas. Like LUNAR, its grammar rules are motivated by linguistic research and, accordingly, can be applied to domains other than the ones for which they were originally developed.[9] DIAMOND/DIAGRAM has four major components: a parser that

---

[9] Systems along these lines have also been developed by Landsbergen (1976), Petrick (1978), Sager (1981), and Thompson and Thompson (1975, 1978). As with LIFER, DIAMOND/DIAGRAM is chosen for illustration here because it is used in one of the projects described in the research section at the end of this chapter.

builds up more complex structures from words and phrases, a general grammar of English, a set of basic semantic operators that actually build semantic structures in the domain, and a set of translators that relate the structural descriptions produced by the parser to the semantic operators. The parser, the grammar, and the translators are all largely domain-independent. In a new application it is necessary to add new vocabulary items and to produce new semantic operators if the domain entails a new kind of representation.

Current natural-language interface systems of both kinds succeed reasonably well in providing analyses of the literal meaning of a request or command. That is, they can relate the particular sequence of words to the database structure through a set of interpretation rules. However, the person using such a system has to be careful about how the request is stated both in terms of grammatical accuracy and to eliminate possible ambiguities. People do make grammatical errors, but often the resulting utterance is easily understood. Similarly, some ambiguities are easily resolved by context, and the mechanisms to accomplish that are easy to program. A variety of procedures are being developed to embody these insights (Kwasny and Sondheimer 1981; Weischedel and Black 1979; Hayes and Mouradian 1980; McKeown 1980).

The present systems also require the user to be familiar with the database and its contents or the associated software program and the commands it uses in order to be able to interpret the system response correctly. Otherwise, he may make assumptions about answers that may not be warranted. For example, in response to the question "How many students failed Computer Science 103?", a negative answer could mean that everyone passed, but it could also reflect the fact that the course was not given. Cooperative systems should inform the user to that effect, and strategies for doing that are being developed (Kaplan 1979; Mays 1980). As interface systems increase in use and acceptance, additional refinements will be requested to support more and more "graceful interactions" (Hayes and Reddy 1979). Similarly, it will be necessary to go far beyond the literal meaning of a request to reflect

the intent behind it and to clarify the larger communicative context in relation to which it is to be understood, objectives that are being addressed in the more basic research efforts of computational linguistics.

b. Understanding Natural-Language Discourse

The objective of facilitating communication in natural language between humans and computers that characterizes work on natural-language interface systems complements the more general concern of computational linguistics with modeling the structure and use of language. Basic research on natural-language understanding, of course, also has its roots in the early developments in parsing and question-answering described above. However, the central goal here is to identify and formalize all the many and complex factors entailed in human communication. In carrying out a program of research, people have been influenced in particular by work in linguistics, psychology and cognitive science, sociology, artificial intelligence and computer science, logic and philosophy. In fact, natural-language understanding has provided a focus for bringing together experts from these various fields (Schank and Nash-Webber 1975; Waltz 1978b).

System building is an integral part of this work, because the complexity of the relations among the constituent elements makes the use of computers extremely useful, if not essential, for evaluating the work. These system implementations also can contribute to practical applications, as illustrated by the DIAMOND/DIAGRAM example, referred to in the previous section. In the long run, these basic research efforts can be expected to provide much more than interface capabilities. They will be particularly important for interacting with the kinds of textual databases we have been discussing in this chapter.

Structural information about the relations among the constituents of a sentence is essential for understanding, but it is not possible to resolve syntactic issues without recourse to semantics. The decision as to whether a sentence is well-formed cannot be made without

some assessment of its meaning. Thus Chomsky's famous example, "Colorless green ideas sleep furiously," contains a valid sequence of grammatical categories, but it can be interpreted only metaphorically at best. Semantic considerations begin at the lexical level as attempts are made to establish a correspondence between words and objects, properties, and events in the world. But an understanding of how words are combined into phrases and clauses requires compositional rules to establish their relations to each other in the context of an appropriate semantic interpretation. For example, in the following sequence of words, consider how the meaning changes as each word is added successively to form a larger phrase: "graduate student apartment rental application."

It has become increasingly clear that it is necessary to incorporate into a computational model of language understanding not only general relations between words, phrases, and sentences and the world, but, in addition, specific knowledge about a particular domain or context of application. Participating constructively in a conversation or reading an article about medicine or politics requires a substantial familiarity with the subject matter. And, of course, the more technical an area of specialization and the more subtle the distinctions to be made, the more refined this information must be.

The formalization and representation of knowledge will be discussed more fully in the next section, since it is one of the major areas of research in artificial intelligence. However, it is useful here to identify some of the approaches being taken. The simplest procedure would be to store attributes and values as tables in a formatted database (Codd 1970), as illustrated in the discussion of natural-language interface systems. Logical representations, especially the use of the predicate calculus (Nilsson 1980) are particularly valuable, because there are formal procedures for making inferences. Semantic networks, in which objects and situations are represented as nodes and the relations among them as arcs, provide a perspicuous display of the structure of knowledge in a particular domain (see the

collection of papers edited by Findler 1979 and particularly the survey by Brachman 1979). Networks are helpful in making the links between language form and information content explicit (Hendrix 1979). Another approach is to express the relations as sets of rules or productions (Newell and Simon 1972) or as frames and scripts that embody the complex structure of objects and events in a given context, for example, activities associated with restaurants or making travel reservations (Minsky 1975; Schank and Abelson 1977; Bobrow and Winograd 1977).

The complexity of understanding human discourse is underscored by yet other conditions that need to be identified. Participating in a dialogue requires a means of capturing the content of previous utterances. Viewed in its simplest form, the words "Yes", "No", or "Seven" as answers to a question are only meaningful when that question is known. Knowledge of the dialogue context is also required to clarify the use of pronouns or definite determiners--"the house", where a specific one is intended; and elliptical expressions--given the question, "Do you have time for a meeting today?", "Tomorrow?" is easy to understand (cf. Webber 1978; Sidner 1979).

To clarify other ambiguities may require reference to a pragmatic context established by previous discourse and by knowledge of changes in the environment that can be expected to occur in the course of a dialogue. Thus, the question, "What bolts are used to fasten it?" becomes understandable (and answerable) following the directive, "Mount the pump on the platform" in the context of an assembly operation in which a series of substeps in the operation are clearly identifiable. Grosz (1977, 1981), Walker (1978), and Robinson et al. (1980) discuss the relevant issues and describe systems designed to participate in task-oriented dialogues.

Complex as the procedures for analyzing syntax, semantics, discourse, and the pragmatics of task context are, they are not sufficient to deal with the more general kinds of communication that take place when a person uses language to achieve a particular objective, and when cooperation with other people is essential to reach

that goal. Here it becomes necessary to understand the plans and goals and the models of the world each participant has, specifically, their respective sets of knowledge and beliefs--both about the world and about each other--which may be inconsistent as well as incomplete, and to be able to reason about them (Allen and Perrault 1980; Cohen 1978; Cohen and Perrault 1979; Moore 1980). Computational linguistics is just beginning to address these problems, and substantial help from research in artificial intelligence will be essential to their solution (Grosz 1979).

The foregoing discussion concentrated on conversational interactions (although one of the participants might be a computer system). However, many of the same considerations hold for the analysis of the processes of communication entailed in reading written documents, although there are distinctive differences as well.[10] In particular, during reading the possibility of accommodation between producer and interpreter is extremely limited. The author of a document has his own goals and plans based on a particular set of knowledge and beliefs. In addition, he has to anticipate for the various classes of potential readers, the range of states of knowledge and beliefs and of goals and plans they might have. However, even the most discerning author cannot write for every reader, and, as indicated above, a document may prove to have a value for some readers unanticipated by the author. While interaction with the author may be precluded, the reader is usually able to review previous parts of the text or even consult other sources of information for clarification.

More significant problems, from the standpoint of computer implementation, result because written texts tend to cover a

---

[10] Rubin (1978) provides an interesting analysis of the differences between oral and written languages in relation to the contexts in which they occur. Actually, working with spoken conversations would require introducing even more complexity into the analysis presented here, for instance, acoustics, phonetics, and the prosodics of stress and intonation. Lea (1980) contains a comprehensive survey of recent research on speech recognition. Walker (1978) presents a system which has addressed the problems entailed in coordinating all the different types of knowledge resources relevant for understanding spoken language.

broader range of grammatical constructions, some introduced for purely stylistic purposes, and because the linguistic devices establishing coherence relations among the text elements may be more complex (Hobbs 1977a).[11] Certainly computer-based grammars, although increasing in scope and complexity (e.g., Woods et al. 1972; Thompson and Thompson 1975, 1978; Robinson 1980; Sager 1981) are not necessarily complete even for the range of potential conversations they are intended to cover. More generally, it is appropriate to observe that no complete grammar of English has ever been written, if indeed it can be, language use being as variable as it is. The classic descriptive grammars, like Jespersen's (1914-1929), were reasonably comprehensive, but none of them was formalized in a way that would allow direct computer implementation, much less any systematic way of assessing their completeness. In any case, these grammars are limited to characterizing the range of acceptable sentences, with little attention to discourse structure.

There have been a few attempts within computational linguistics to analyze texts in order to derive their structure. Sager (1981) and her colleagues, working with scientific articles in pharmacy and medical discharge summaries, have developed procedures to extract information from a document and convert it into tabular form. It is possible then to query the file structure and derive answers to specific questions (Grishman and Hirschman 1978). Schank and his students have done extensive work in processing stories and even newspaper articles (see Schank et al. 1980, for a summary of this work). The information they contain is converted into a complex "conceptual dependency" structure, on the basis of which questions can be answered and paraphrases and summaries generated. The systems developed by Sager and Schank are still experimental prototypes, but they demonstrate progress toward the goal of being able to process a text automatically to derive a representation of its content.

---

[11] As a compensating factor, deviations from grammaticality are less frequent in written language, and the author usually does have the opportunity to consider more deliberately how to express what he wants to convey, although he may not always take advantage of it.

Work on machine translation, which obviously has to draw on the same linguistic and computational resources, has not resulted in any more practical procedures for handling text (see Hutchins 1978). Many current systems provide little more than dictionary lookup and sentence-for-sentence conversion, although there has been some success in processing short texts, for example, weather reports, where the language is stereotyped and the semantics narrowly constrained (Chandioux 1976). Given this state of affairs, there has been an increasing emphasis on machine-aided translation and on computer aids to translators (Zachary 1979), not only to provide more practical help, but in recognition of the value of being able to observe translators at work and learn from their actual practice. The results of these observations could contribute both to translation and to language understanding more generally.[12]

It is clear from this discussion of the problems of understanding natural-language discourse that the solutions will be complex. However, the results of current research can be expected to contribute both in increasing the sophistication of language interface systems and in increasing our knowledge of the structure of language and the processes entailed in its use.

3. Artificial Intelligence

The process of synthesis and interpretation requires intelligent behavior. Formalizing the principles that underlie intelligent behavior and developing computer facilities capable both of supporting and actually demonstrating intelligence constitute the program of research undertaken by the field of artificial intelligence. This distinction between formalizing principles and developing computer facilities is not simply a contrast between science and engineering, although that does constitute one dimension for characterizing work in the field. Rather, the computer implementations play a key role in the development and testing of theoretical formulations. It may be

---

[12] Karlgren, Walker, and Kay (in preparation) provide a comprehensive examination and evaluation of the role of computers in translation.

appropriate to remark that, without having computers to cope with the complexity of the subject matter, it would be extremely difficult, if not impossible, to work on these problems at all. Within artificial intelligence, there is a tension between approaches that base their work directly on models of human psychological processes and those that are motivated more by logic and an interest in algorithms and abstract procedures for heuristic search.

Research in artificial intelligence can be characterized by the kind of activity or area of behavior studied or by the basic concepts and techniques that reflect underlying mechanisms. In the first case, it is appropriate to refer to vision and image analysis, language and speech understanding, robotics, knowledge-based systems, automatic programming and program synthesis, distributed data management, and game playing. Work in one of these areas would usually be described in relation to a system that performed at least some of the behavior associated with the activity. Considered in relation to the concepts and techniques, artificial intelligence is concerned with issues of representation and modeling, inference and common sense reasoning, knowledge acquisition and use, heuristic search procedures and system control structures. Research on concepts and techniques is often motivated by the desire to formalize abstract principles. However, as noted above, these principles are almost always developed and tested in the context of a system implementation.[13]

Language and speech understanding and knowledge-based expert systems are the two areas of artificial intelligence research most relevant for consideration in the context of knowledge synthesis and interpretation. Since much of the material on language has been considered in the previous section, the discussion here will concentrate on knowledge-based systems. This focus makes it easier to illustrate the concepts and techniques most appropriate for consideration, although at some risk of oversimplification.[14]

[13] Two recent texts in artificial intelligence, both of which can be recommended as introductions to the field, reflect this contrast: Winston (1977) builds his around particular systems; Nilsson (1980) emphasizes underlying principles.

Work on knowledge-based systems began in the mid-1960's in part in reaction to the failure of attempts within artificial intelligence to develop general, computer-based, problem-solving strategies that would be applicable in any situation. But it also reflected the realization that complex decisions required specific kinds of knowledge, often in large amounts and with subtle interrelationships, and that computers might be some help in sorting out the relevant information. The task addressed in what was probably the first knowledge-based system, DENDRAL, was inferring the chemical structure of organic molecules (Feigenbaum et al. 1971). The inferences were based on rules developed by chemists doing research on mass spectroscopy and nuclear magnetic resonance that related data from their experiments to specific atomic configurations. Using these rules and predicting spectrogram results according to specifiable constraints on their interactions, the system was able to test, refine, and even extend knowledge in the field (Feigenbaum 1978/1979).

Knowledge-based systems have now been developed in a number of areas, although most still are to be considered in the research stage. Medical diagnosis is well represented by systems that help physicians identify bacterial infections and prescribe drugs for their treatment (Shortliffe 1976) that provide differential diagnosis in internal medicine, distinguishing on the basis of symptoms and signs among hundreds of diseases (Pople et al. 1975; Pople 1977); and that model the relation between clinical observations and pathophysiological processes in the eye occurring in glaucoma (Weiss et al. 1978); to mention only a few current efforts.[15] Another system that is particularly well

[14] Smith (1980), in a comprehensive review of artificial intelligence research in relation to its application to information systems, singles out pattern recognition, representation, problem solving, and learning as the primary focal points. Her treatment of these issues from a more general perspective provides a useful complement to the approach taken here.

[15] Shortliffe, Buchanan, and Feigenbaum (1979) review the full range of work in medical decision making in a way that shows clearly how approaches based on artificial intelligence techniques differ from other computer-based clinical aids.

developed incorporates geological knowledge and advises exploration geologists on the mineral deposits likely to be present in a particular location (Hart et al. 1978; Duda et al. 1979). In spite of the relatively early stage of work on these kinds of systems, the accomplishments of recent research have led to the characterization of this approach as "knowledge engineering" (Feigenbaum 1978/1979), and to predictions that it will revolutionize the way people in the fields to which it is applied will work. Reflecting the fact that there are similar procedures in the different systems, a facility has been designed to make it easier to build new ones (Nii and Aiello 1979).

The knowledge that is incorporated into a knowledge-based system is derived from the experiences and practices of "experts." However, much of the material required cannot be described spontaneously by the scientist or professional even though it is used regularly as the basis for making judgments. It often does not constitute a formal element conveyed explicitly as part of the educational process in a discipline. Accordingly, one of the first problems in the development of such systems is how to represent the relevant knowledge, that is, how to describe it so that it can be stored in computational form. Once an adequate system for representation has been established, the next problem is how to make inferences about the relations among knowledge elements and to apply the results in a specific instance. In addition, for a system to be trusted, it is essential that the inferences and their results be made perspicuous--in effect that they explicate the kinds of decisions an expert would make in the same situation. As a result, an explanation component must be included to describe the actions taken by the system, justify their ordering, and substantiate the conclusion. Finally, in order to extend the system it is necessary to develop procedures that make it possible to extract new knowledge systematically from the human expert. This set of problems corresponds in good measure to the list of artificial intelligence concepts and techniques mentioned earlier. In the following discussion, representation is stressed, since that is the basic question.

Although representation is a central issue in every area of artificial intelligence research, there is no consensus on a single approach. In fact, a recent survey of knowledge representation revealed a large variety of different views on the subject (Brachman and Smith 1980). This finding is not surprising, considering that philosophers have been arguing about the problem for over 2500 years. However, people do not even agree about whether what is being represented is the world, some symbolic system (i.e., the embodiment of "knowledge"), or the practice engaged in when one attempts to assign some symbol to something called knowledge. While it is not possible to resolve these questions on the philosophical level, fortunately that is not necessary for actually building systems, although it will affect the degree to which such systems can be generalized and extended. In any case, for the present purposes it will be sufficient to characterize the major ways in which artificial-intelligence systems are addressing the representation problem.

Four approaches to representation are particularly worth describing: the predicate calculus, production systems, semantic networks, and frames. In the predicate calculus, representations are formed by grouping elements consisting of constants, variables, functions, and predicates, into formulas through sets of connectives (like "and," "or," "if-then") and quantifiers (indications of relative scope corresponding, for example, to assertions that a statement is universally true or that something exists) (see Green 1969; McCarthy and Hayes 1969; Nilsson 1980, Chapter 4). The predicate calculus has both a well-defined syntax, that is, a formal set of rules for establishing the well-formedness of the formulas, and a well-defined semantics, that is, the truth or falsity of the formulas can be decided by establishing correspondences between the elements and the part of the world that constitutes a domain of application and by carrying out the appropriate calculations. Arguments for the predicate calculus are that it is domain independent and logically complete and that there are mechanical proof procedures for deriving conclusions from a given set of premises. Arguments against it are that it is not possible to capture special

relationships that exist in a given domain, that it cannot tolerate inconsistencies--for example, those that would be involved by incorporating into a system different sets of beliefs, that it is difficult to understand the logical formalisms, and that the proof procedures are too long and complex to be practical.

In a production system approach, knowledge is represented as a set of rules that relate a specific state or condition or antecedent to a particular action or consequent (see Newell and Simon 1972; Davis and King 1976; Waterman and Hayes-Roth 1978). The rules can be interpreted in either direction: if a condition is present, the corresponding action can be initiated; if an action is desired, the corresponding condition should either be established or looked for. Once a database of such rules has been developed, it is possible to apply them systematically in a given context, in effect generating and testing hypotheses until one that applies is found. Since the rules embody specific conditions and actions, they can be used to encode information specific to a particular topic. Moreover, the database can be augmented as new rules are developed. The major problems in production systems seem to arise in establishing relations that show different kinds of dependencies among the rules in a set and in establishing orderings for the rules that provide optimum efficiency in analyzing a complex situation.

Semantic networks consist of nodes connected by arcs (see Quillian 1968; Simmons 1973; Woods 1975; Findler 1979; and Hendrix 1979). The nodes correspond to physical objects, situations, events, or sets; the arcs specify any of a variety of relations among the nodes. Thus the networks can represent factual elements at varying levels of complexity as well as the associative relations among them, both direct and indirect. One distinctive feature of a semantic network is that it shows explicitly the interconnectedness among objects and relations. By introducing additional structures into a network, it is possible to incorporate hierarchical information and logical connectives and to establish the appropriate scope for quantifiers. These additions, however, reduce the perspicuity of the network and increase the complexity of carrying out operations on it.

Frames are data structures that constitute prototypical objects or events (Minsky 1975; Bobrow and Winograd 1977; Schank and Abelson 1977). Each frame contains a set of slots that relate to distinguishable aspects of the object or event. A person frame might have a name, age, address, marital status; a trip frame might have an origin, a destination, times for departure and arrival, a mode of transportation; an event frame might register who, what, where, when. Frames may be linked together in various ways to facilitate establishing relations among them; for example, a slot may contain a pointer to another frame. It is also possible to associate a procedure with a particular slot so that some action is initiated whenever that frame is called on. A frame system can be used to embody complex descriptions of events, but problems arise in deciding what frame applies in a given situation, in determining when and how to shift from one frame to another, and in reasoning with such data structures.

Deciding which kind of representation to use is certainly affected by personal preference, educational experiences, and sociological factors--there are fads and fashions in artificial intelligence as in any field. However, there are a number of criteria that are often invoked (see Winograd 1975): flexibility and economy--being able to use a particular knowledge element easily in a variety of ways; understandability and learnability--being able to see what elements are in a system and how they are related to each other; accessibility and modifiability--being able to locate a given element and to modify it without difficulty. Balanced against these criteria are ones that reflect on the complexity and selectivity of the relations that actually hold among the knowledge elements; it may be desirable to build some of these into the system and to use heuristics, special strategies that are unique to that area. In addition, there also are special characteristics of the subject matter that must be taken into consideration. For example, expert systems in both medicine and geology must deal with probabilities; a given set of observations and laboratory tests do not inevitably lead to a single conclusion. Moreover, in those fields a number of different causal or etiological factors may

contribute to the observed result. For geology, "plausible" as well as logical and contextual relations are encoded in an inference network to link field evidence and geological hypotheses, using Bayesian statistics (Duda et al. 1976).

Once the body of knowledge relevant for a knowledge-based system is represented and stored, using it requires finding elements that correspond to a search specification, either wholly or in part. If the match is not complete, it is necessary to perform inferences that establish relations among knowledge elements. It should be apparent from the discussion of representation that the procedures for matching and inference are not independent of the form of that representation. As noted, for the predicate calculus, there are well established formal proof procedures (Loveland 1978; Robinson 1979). Inference procedures also have been developed for production systems (Hayes-Roth et al. 1978), semantic networks (Fikes and Hendrix 1977), and frame systems (Bobrow and Winograd 1977; Schank 1979). Of particular interest for future work in expert systems are recent attempts to formalize common-sense reasoning, leading to what are coming to be called non-monotonic logics (McDermott and Doyle 1980).[16] In such logics, inferences can be made based on incomplete information, and changes in assumptions (axioms) can invalidate previous beliefs (theorems).

Other topics related to knowledge-based systems can be mentioned more briefly. It is essential that such systems be able to provide the user with explanations that show how a particular conclusion has been derived. Explanations are, of course, complementary to inferences. However, it is critical to be able to describe, in a form that is clear and understandable to the user, exactly what the successive steps in the process are. The medical systems referred to above have been especially concerned with explanation and reasonably successful in accomplishing it.

---

[16] A double issue of the journal Artificial Intelligence (Volume 13, Numbers 1 and 2; April 1980) contains seven papers devoted to this theme.

The knowledge bases for most of these systems have been elicited from experts through personal interactions with the system developers. The development of computer-based procedures to assist in the acquisition of new knowledge is beginning to receive more attention. One such system, used in conjunction with the geology exploration system, simplifies the process of building the complex network structures required (Reboh 1981). Another, used in conjunction with the medical system for diagnosing bacterial infections referred to earlier, interacts directly with the expert to add new rules or augment older ones, building on its understanding of its own knowledge structures (Davis 1979). Perhaps the most ambitious of these systems is intended to acquire knowledge through tutorial dialogues in English, translating the knowledge directly into a logical representation (Haas and Hendrix 1980).

The objective in all of these knowledge-based systems is to synthesize a body of knowledge for a particular application. Since the enterprise is so new, there has not been enough experience with them yet to appreciate how this embodiment differs from that in more traditional forms, like textbooks or manuals. In particular, we do not really know how people will use them to find the information they need. Accordingly, it is particularly appropriate that they be evaluated in the context of a program directed towards determining how people actually organize and use information.

C. AN APPROACH TO THE ORGANIZATION AND USE OF INFORMATION

The remainder of the chapter describes a research program specifically designed to study the organization and use of information. The research strategy entails the development of systems designed to provide natural-language access to both data and text files for scientific and professional work. The initial efforts are being carried out in the fields of medicine and law, but the results are expected to have much broader applicability. The intent is to provide physicians and lawyers (and, more generally, any professional involved in knowledge

synthesis and interpretation) with tools that allow them to formulate requests for information, essentially in the form that they come to think about them in the course of their work, and to interact with the system in a dialogue as they progressively refine and clarify their hypotheses and their understanding by revising their requests based on responses to the material retrieved.

A major objective of the program is to gather substantive data about the nature of problem formulation, what uncertainties motivate the requests people address to the system, how these uncertainties relate to their perceptions of the conceptual structure of the field, and how the data or textual materials stored in the system and displayed in response to a request resolve the uncertainties that motivated it. Our understanding of this process will be sharpened by systematically modifying the ways in which the material is organized within the system as we see how people use it. In addition, it will be desirable to let the user build up intermediate knowledge structures on the basis of his progressive insights into relationships among data elements.

The systems we are designing and testing build directly on the results of research in information science, computational linguistics, and artificial intelligence described in the preceding sections. Moreover, they will be able to take advantage of, as well as contribute to, new developments. In effect this research program is at the intersection of these fields: it reflects the interest of information science in the nature of information and the development of information-retrieval systems; it incorporates computational linguistic techniques to allow communication in natural language between the user and the system; and it leads to the creation of knowledge structures of the kind developed in artificial intelligence that both support the use of the system and increase our understanding of the processes involved.

Our approach to the study of the organization and use of information is based on the premise that an information-retrieval system, properly configured and easy to modify and augment, can constitute an environment for carrying out productive research on this

problem. A primary requirement for such a research program is that the people actually using the system have a need for information and that they can determine, on the basis of the data retrieved, whether that need is satisfied. Accordingly, we are addressing the end user and not an information intermediary. To encourage this class of users, it is important to make the system easy and natural to use. This requirement is reflected both in the way people view the data stored and in the manner in which they interact with those data.

The user should be able to think of the primary data in the system in the way that training and experience in a particular discipline have led him to consider them. That is, the system should contain material in substantially the form that it appears in disciplinary resources. It is essential to include the full text of a document, preserving as much of the structural relations among the text elements as possible, including, in particular, the sequential and subordination relations reflected in chapter, section, paragraph, and sentence organization.

Similarly, we want to allow a query or hypothesis to be expressed as far as possible in the form the user comes to think about it. We believe that this is most often in natural-language form, but that term, as used here, includes the jargon and the shorthand conventions that specialists in a field come to employ. For more complex queries, the initial formulation is rarely the most appropriate one, because the person usually is not clear about just what he or she wants to know, and because it is not clear exactly how the data that might constitute answers are embodied in the system. Consequently, it is essential for the user to be able to engage in a dialogue with the system in order to refine the query or hypothesis through successive interactions.

The representation issue, how to characterize the content of the query and the content of the data and how to establish procedures for relating them to each other, is a central concern of the entire research program. On the simplest level, the mapping from query to data might involve matching terms that occur in each. Identifying more complex relations entails being able to specify the propositional structures of

both queries and data elements. The system structure we are proposing will allow exploring the implications of the range of possible alternatives and evaluating their effectiveness under particular conditions. A person may not be able to find relevant data because of problems in the analysis of the query, in the representations of the content of the data, or in the mapping functions themselves. Some of these problems may be resolved by the presence of alternative representations or different mapping functions; some through browsing or random exploration by the user; others will require intervention by the system staff and even by experts in the subject matter represented in the database.

When one or more sets of data are determined to be relevant, the user must be able to enter into the database his evaluation of each set, specifically in relation to the query or hypothesis in its refined form and implicitly or explicitly in relation to the intermediate representations and mapping functions that led to those data. These evaluations are to be stored so that subsequent users, when they enter similar queries, can be directed to the evaluations of their predecessors, as well as to the original source data. In this way, the database can accumulate the experience of its users. One of the most significant failings of current information-retrieval systems is that each user interacts with them as if they had never been used before.

The capabilities of the system also must support more direct interactions among the users: there can be dialogues between users about the data as well as between a user and the data. It will also be possible to enter into the system as primary data elements those documents created by the users as a result of their retrievals. In this way, the system comes to be a continuing resource for synthesis and interpretation in a particular field.

Of course, not every source is equally credible, so it is critical that each item in the database be identified as to authorship. This requirement applies to metatext as well as to primary text, and must include programs as well as people. Another important feature is to be

able to characterize the degree of confidence or certainty associated with each item. Tentative hypotheses, labeled as such, can provide the basis for new directions, and are particularly valuable when they represent the convergence of different perspectives.

It is possible, of course, that even though people formulate similar queries, they would not find the same material relevant. Information is not intrinsic in the data, as noted in the discussion on information science above, even though the producer of a document may intend that it be interpreted in a particular way. Rather, the value for the user is a function of the way in which the data satisfy that person's need. Furthermore, the query, at least in its original formulation, may not reflect that need accurately.[17] Therefore, it is essential that careful analyses of the transcripts of retrieval sessions be carried out in conjunction with in-depth interviews of the people who made them. The insights derived from these studies and from an evaluation of the system over time as changes are made to increase its effectiveness should contribute significantly to our understanding of the organization and use of information and, thus, to the concept of information itself.

We believe that the system proposed here has general features that are independent of specific fields of knowledge and specific areas of application. The particular kinds of data used and structural features of the knowledge that represents their content will vary from field to field, but at what level of specificity and with what implications is not clear and will require extensive study to determine. Similarly, although we assume that the language used for interacting with the system will have general properties, it will vary in lexicon for different applications and probably in certain types of syntactic

---

[17] Belkin's (1978, 1980) concept of an information need as an anomalous state of knowledge is quite relevant here. A person may recognize that he needs to know something more about an area but not be able to specify it more precisely. Belkin and his colleagues (1979; also see Oddy 1977) have been studying the relation between a user's expression of needs and representations of the material that seem to satisfy them to develop some insights into this problem.

constructions. These differences in language and representation will be central topics for future research.

Our strategy for carrying out this program of research entails implementing prototype systems that explore selectively certain of the capabilities that have been discussed. Accordingly, the following three sections describe our current efforts. The first, "providing natural-language access to data," focuses on the utility of a natural-language interface for retrieving formatted information from a database in the context of actual use by a person familiar with the structure of the file. The second, "representing knowledge in texts," concentrates on the development of procedures for analyzing the propositional content of text passages so that natural-language queries can provide selective access. The third, "facilitating access to information and communication among users," is directed toward establishing a more general system structure in which the primary data elements (e.g., text passages), queries addressed to retrieve them, and comments about them are all handled in the same way, effectively as messages.

1. Providing Natural-Language Access to Data

The immediate objective of this project is the continued development and testing of MEDINQUIRY, a system that allows physicians to retrieve data about patients through requests formulated in natural language. It was designed to take advantage of the availability of procedures for retrieving data from formatted files and recent developments in natural-language interface technology. The decision to choose a medical application reflected specific interests of the staff members involved, but it also recognized the existence of a large body of ongoing research in artificial intelligence in medicine. Funded by the National Cancer Institute[18], the project is a collaborative effort involving SRI International, the University of California (San Francisco), the University of Pennsylvania, and Martin Epstein of the National Library of Medicine.

---

[18] Grant No. 1 R01 CA26655.

The prototype MEDINQUIRY system (Epstein and Walker 1978; Epstein 1980) contained a database of information on 130 patients with melanoma, a malignant skin cancer. The information was taken from the patient records, and included personal and family histories, physical examinations, clinical and histological studies of the tumor, diagnosis, therapy, and various kinds of follow-up material, 156 attributes in all. LIFER, referred to in the section on Computational Linguistics, was used to create the natural-language interface. On the basis of a thorough review of the literature on melanoma and consultations with physicians who specialize in the disease, the kinds of requests that seemed to be of particular interest were identified, a grammar that could analyze them was written, and the retrieval functions that return relevant data were established. We are currently engaged in expanding the database which will eventually contain information on more than 1500 patients and in augmenting the system to accommodate more attributes.

The system is intended to satisfy two major kinds of uses by the physician: patient management--how treatment of a given patient should be influenced by the results achieved with similar patients; and basic research into the clinical course of the disease--testing hypotheses about attributes relating characteristics of the patient and his treatment to survival. The physician can request data on selected attributes and their values for a particular patient or for groups of patients satisfying certain characteristics. He can ask for simple calculations to be performed, identifying frequencies, averages, ranges, and the like, and allowing information to be tabulated in various ways. It is possible to browse through the database, identifying and studying relationships among patient attributes, and these relationships can be correlated with prognosis and outcome.

Examples of requests that can be processed by MEDINQUIRY are:

How many patients in the melanoma database?
What are the classes of attributes?
List the items associated with pathology results.
Display the path results for patient s-72-002.
Count the number of individuals who had lymph node procedures.
In how many patients was level 5 disease seen?
List the age, sex, and tumor thickness for people with superficial

spreading melanoma level 5.
Stratified by site of the primary, how many people had ssm followed by histologic recurrence within 1 year of initial therapy?

The system supports dialogue interactions; the user can follow a line of inquiry to test a particular hypothesis by entering a sequence of requests that depend on each other. Phrases rather than complete sentences can be used, where the meaning of a phrase is interpreted based on an analysis of a prior request. It also is possible to define new patterns for phrases and sentences and to store, in a generalized form, sets of requests that might be used frequently.

Once the number of patients in the system is sufficient to be able to provide reliable medical insights, we will begin analyzing and evaluating how physicians use the system. Transcripts containing complete records of their interactions are gathered automatically by the system. However, it will be necessary to discuss each with the particular individual involved to determine the rationale underlying the sequence of requests that he made. While the experiences of the users will guide further modifications of the system, the primary interest will be in understanding how medical decisions are made both for patient management and clinical research. In particular, we expect to gain insights into the process of hypothesis formation done by medical experts, the strategies they use in establishing relationships among variables, and the kinds of knowledge structures they find useful in the process. Eventually, it may prove possible to formalize these relationships and the knowledge structures that are developed and build them back into the system, effectively as a "knowledge base" that would serve as a model of relations in the database.

In the context of the overall program, this project establishes a baseline. The database contains, appropriately formatted, all of the information about the patients that is currently considered relevant. The physicians who constitute the initial user population are acknowledged specialists in melanoma. They already have hypotheses they want to test. The natural-language interface capabilities allow them to formulate requests in English. Consequently, we can now begin to

evaluate how they organize and use the information our system can provide.

2. Representing Knowledge in Texts

The long-range goal of this second project is to provide scientists and other professionals with access to textual materials through natural-language requests, paralleling the work on natural-language access to data discussed in the previous section. However, our understanding of how to structure text files is nowhere near as advanced as that of formatted file management. The existing procedures for locating a particular text passage in a document are both awkward to use and grossly insensitive. Accordingly, our primary focus in this work is on strategies for representing the information contained in a document. To test our results on a continuing basis, we are embedding the texts and their representations in a system that allows requests to be formulated in English, but we are much further from being able to actually allow anyone to use the system to satisfy a real need in this project than in MEDINQUIRY.

The choice of a medical context this time was motivated primarily by the availability of a computer-based medical textbook being developed at the National Library of Medicine (Bernstein et al. 1980). This textbook, called the Hepatitis Knowledge Base (HKB), is intended as the first in a series of documents that will eventually span the field of medicine and allow physicians and other health professionals to have available, in easily accessible form, the results of current research in their field.

Since the HKB was designed as a computerized textbook with selective retrieval as a primary goal, it is systematically organized in a hierarchical structure. In its current implementation on a minicomputer, the user locates relevant information by reading through a detailed table of contents and selecting particular passages corresponding to the headings listed there. Our intent, by annotating text passages on the basis of their information content and providing

natural-language access, is both to increase the efficiency of the retrieval process and to be able to understand better how people actually work with these materials. In this sense, the work on this project parallels the one just described. However, using the HKB, which is the focus for a variety of studies, will make it possible to compare the procedures we develop with other approaches, thus providing another basis for assessing the results.

The immediate objective of the project, which is supported by the National Library of Medicine,[19] is to develop a prototype system that will contain two major components: (1) a natural-language interface capable of analyzing a wide variety of English grammatical constructions; and (2) text access procedures based on representations of both the propositional content of summaries of passages in the text and the hierarchical structure of the text itself (Walker and Hobbs 1981).

The natural-language interface is being provided by the DIAMOND/DIAGRAM system, which, like LIFER has also been discussed in the section on Computational Linguistics. It consists of an annotated phrase-structure parser with a linguistically motivated general grammar of English (cf. Robinson et al. 1980 Robinson 1980). The parser produces phrase-structure trees that are translated into semantic operators to establish an underlying semantic representation; other translator functions make it possible to check on discourse context. The parser, the grammar, and the translators are largely domain-independent. Application to the HKB requires entering the relevant vocabulary and analyzing the subject matter to produce the appropriate semantic representations.

The text access procedures are being developed on the basis of a predicate/argument analysis of the content of text passages in the HKB. The informational content of each passage is expressed as a conjunction of propositions. The overall text structure for a document consists of the set of conjunctions corresponding to the constituent

---

[19] Grant No. 1 R01 LM03611.

passages, together with representations of their organization into a hierarchy and with pointers to the associated passages. Note that, at this stage of our work, the text representations are being developed manually; our concern is to establish their utility. Having done that, future efforts will be devoted to automating procedures that can make those assignments.

As indicated, the establishment of this text structure constitutes the primary focus of our initial efforts on the project. However, we will be developing other aspects of the prototype system in parallel so that we can test the concepts on a continuing basis. In particular, the system must be able to analyze a request and translate it into propositional form, to resolve anaphoric expressions, to reduce the request to a canonical form via synonymy and other inferential relations, and to match the request against the text structure, returning to the user a relevant passage or set of passages from the text.

In the context of the overall research program, the initial efforts on this project are intended to establish procedures for working with text that parallel those now available for working with formatted data. However, where in our first project the physicians could be expected to know the attributes associated with patient records, it is not likely, even with medical guidance, that our initial representations will be equally "natural" for our users. Consequently, we anticipate modifying our assignments systematically on the basis of experience. Having the primary text as the material actually displayed provides us with an opportunity to determine exactly what materials constitute information appropriate for a user's needs.

3. Facilitating Access to Information and Communication Among Users

The two projects previously discussed address two key problems in the development of this program of research on the organization and use of knowledge: investigating the utility of natural-language access

to a formatted database by professionals familiar with the information it contains; and developing procedures for analyzing the propositional content of text passages. This third project addresses the more general issue of establishing a comprehensive system framework in which this other work can be embedded. As such it considers more fully the set of problems delineated in the introduction to this section in which we described the approach we are taking. This third project is both more simple and more complex than the other two. It is more simple in that it can operate with a lower level of sophistication in the procedures for accessing information. It is more complex in that the objective is to develop a general facility within which the procedures developed in the other two projects can eventually be accommodated.

The activity in which we are engaged reflects a new concept for information retrieval design, called "Polytext" (Karlgren and Walker 1980). It is being developed by SRI jointly with Hans Karlgren and the Kval Institute for Information Science in Stockholm, Sweden. There are three distinctive features of the approach that respond directly to concerns already expressed in our earlier discussion.

The first is that Polytext provides a basis for incorporating into a database the cumulative experience that users have had with its contents. Thus, each time a person enters a request into the system, he can access (1) the original source documents (which may be texts, sentences or paragraphs within a text, thesaurus lists, etc.);[20] (2) citation data, index labels, summaries, abstracts, and other representations of the content of those documents, identified as to their source; and (3) pointers to previous requests that are similar to his, to evaluations that other users have made of the responses provided by Polytext to those requests, and even to more general evaluations of the source documents or the various content representations of the documents themselves.

---

[20] Although we believe the Polytext concept is broad enough to encompass formatted files as well as text, in the current work we are concentrating exclusively on textual materials. Consequently, the following discussion will reflect this narrower focus.

The second feature of Polytext is that it allows a user without computer experience to interact with the system in relatively natural ways. By natural, we mean that people who are specialists in a particular field of inquiry are able to address the system in the same way they think about their problems and communicate with their colleagues. Consequently, Polytext provides for a dialogue between the system and the user in the refinement of search requests and in the evaluation both of intermediate and final results. In particular it is possible to formulate a request in natural language, and procedures are available that allow sequences of requests to build on each other successively. Similarly, as the system returns information in response to those queries, facilities are being developed so that the user can record his assessment of each item. These assessments, themselves evaluated as a whole by the user at the end of an inquiry session, will be incorporated into the system database.

The third distinctive feature is the accommodation of a variety of algorithms and strategies--manual and automatic--both for accessing and for processing texts so that comparative evaluations of their relative effectiveness can be made. The procedure responsible for the response will be identified explicitly, and the user will be able to select, and, in principle, to modify them, although that might require a higher level of sophistication than that available to most users.

For textual material, the source documents to be included are stored in their entirety. The primary response to a request usually will be a passage or set of passages that is appropriate, based on an analysis of the request and a comparison of the results of that analysis with the sets of representations of the text passages. There will be successive improvements in the sophistication of the representations, and hence of the precision of the response, guided by experiences in the use of the system. However, the intent is to allow the user, viewing the text retrieved, to make his own evaluation of its adequacy with respect to his information needs.

The system must be dynamic, that is, material will be continually added to it, both in the form of new primary texts and in the different kinds of metatexts: that is, the representations associated with primary texts, the requests addressed to the system, and various kinds of evaluations. The basic mechanism required for the system, therefore, is that it be able to handle all of these materials, essentially on the same level and in the same manner. The model we find most appropriate here is to consider both primary and metatexts as messages. The basic operation, then, is manipulating messages and the structure we are proposing is analogous to that used for message switching and computer conferencing systems.[21] However, in the course of message manipulation, we must provide procedures for analyzing texts, both structurally and semantically, for linking messages so that they are related in dialogues, for performing inferences, and for carrying out searching and matching operations.

To satisfy the requirements that have been described, we have developed a set of design specifications. The way the system handles messages should result in the following pairs of "users" being treated in the same way: (1) authors of primary documents and those who comment on them, index them, or assign some other kind of content representation to them; (2) information providers and clients--any user may introduce new messages at any time; (3) people and machines--programs as well as human users of the system may read and supply messages.

We have attempted to keep the basic Polytext software as simple as possible. Therefore, intelligent and short-lived modules are kept outside as programs using the system rather than as parts of it. Thus, text analyzers (machine or human) may take one message at a time, interpret it, and report the result as a new message, which has the analyzed message as its topic and the recoding in some meta-language as its comment. Naturally, the lexicon for the system would be stored in message form, and the programs could use other messages for information in the course of their analyses.

---

[21] See the chapter on computer-based communication systems in this book by Johansen and Vian.

Stated in the most basic terms, the system functions can be characterized as follows: (1) insert a message; (2) delete a message; (3) read a message explicitly specified by its identifier; (4) supply the identifier(s) of a message (or the messages) linked to a given message; (5) link messages as prescribed by topic-comment information, that is, indications of dependency relations, and by the pragmatic item, which directs that special functions be performed.

Having set out the design features for Polytext, and recognizing the need for a project of this magnitude to proceed by well-defined steps, the first phase of our research included producing a demonstration model in which some of the basic concepts could be verified (Loef 1980). For our initial work, we selected a short legal document, the "Rules of the Arbitration Institute of the Stockholm Chamber of Commerce" (Stockholm Chamber of Commerce 1977).

Working with Kval, we developed three ways of providing access to the text: using index terms, using the hierarchical structure of the text, and using an analysis of the predicate-argument, or propositional, structure of the text to derive a more detailed model of the information it contains. For each approach, we provided the appropriate interface to a LIFER grammar so that it was actually possible to enter English queries and to retrieve the appropriate passage as a response.

For these pilot studies, we focused on the problems entailed in representation. The text contains twenty rules, each of which has one or more paragraphs or subsections. Four rules deal with the organization of the Arbitration Institute itself, identifying objectives, composition of the governing Board, what constitutes a quorum, voting, and the like. The other sixteen define the arbitration process: specifying how the arbitral tribunal is to be appointed, how a request is filed, the contents of the statements of claim and defense, procedures, voting, awards, and the like.

In the index-term approach, we assigned to each paragraph or subsection one or more index words or phrases, e.g., "arbitrator", "number of arbitrators", "disqualification", "appoint chairman of

tribunal." A relatively simple grammar allows the following kinds of questions to be asked: "What do you know about voting?", "What is said about awards?", "What information is there about the arbitral tribunal?" The system responds with the relevant passage and its rule number and section. Index terms can be added on line as comments to a particular passage.

The hierarchical approach extends the index-term approach by building on structural relationships within the text itself. For example, there are two rules dealing with the request for arbitration; one establishes how it should be submitted by the claimant and what it contains; the other specifies what is to be done with the request by the Arbitration Institute and what the reply by the respondent should contain. Consequently, each of the rules deals with both requests and contents. By taking into account the organization of the text, the request "What should the request for arbitration contain?" can be unambiguously taken to refer to the former. The passage on contents there is immediately dependent on the one specifying request, whereas in the other it is dependent on the one specifying reply.

To provide a clearer linguistic motivation for the assignment of hierarchical structure, we analyzed the text to determine its topic/comment relationships (Kiefer 1979; cf. Halliday 1967; Sgall 1977). This kind of analysis, which constitutes a major focus in contemporary linguistics, illuminates both the functional organization of texts and the semantic structure of sentences. It will be a major topic for subsequent research.

The third approach, the propositional approach, introduces a more complex set of relations among the concepts in the text. It is based on the identification of predicates and their arguments, and it provides for a more precise discrimination based on the direction of relationships. To illustrate, according to the rules for the Arbitration Institute, the Chamber appoints the Board and its chairman; and the Board, in turn, appoints the chairman of the particular tribunal for an arbitration and (under special circumstances) other arbitrators.

Consider the queries:

Who appoints the Board?
Who does the Board appoint?

An index term approach would return the same set of passages for each query. With the propositional approach, "the Board" in the first query is recognized as the object; in the second query, "the Board" is recognized as the subject; different passages are returned for each.

We have implemented a detailed model of the arbitration process, motivated in part by the case grammar analysis developed by Fillmore (1968). It allows us to establish, for a given predicate, such arguments as "who", "what", "for whom", "according to", and "if". To provide the matching of propositional structures required, we employed partitioned semantic nets as our underlying representation (Hendrix 1979) and adapted the inferencing capabilities developed for a system that converts English descriptions of algorithms into programs (Hobbs 1977b). This prototype Polytext system needs to be expanded so that we can test the whole set of procedures already designed. And, of course, we must get people to use it who actually need the information it contains so that we can guide modifications on the basis of their experiences.

In the context of the research program, Polytext constitutes an environment in which the range of issues associated with the organization and use of information can begin to be evaluated. It will be essential to incorporate meaningful capabilities for dialogue interaction and content representation, of the kind being developed in the other two projects, but the establishment of a flexible system structure that can accommodate many users and can accumulate their experiences is critical.

D. CONCLUSIONS

The approach to knowledge synthesis and interpretation taken in this chapter derives from a more general interest in research on what we have called "the organization and use of information." The research program we are engaged in derives from recent developments in information science, computational linguistics, and artificial intelligence. Accordingly, it has been appropriate to describe work in those fields in some detail to provide the necessary context. With that material as background, I presented the objectives and the approach being taken to achieve them as exemplified in three related projects:

> (1) Developing and testing a system that allows physicians who specialize in skin cancer to get information about patients with malignant melanoma for both patient management and clinical research.
>
> (2) Formulating procedures for representing the information content of a computerized textbook on infectious hepatitis so that physicians and other health professionals can find passages relevant to a particular problem.
>
> (3) Establishing a more general system framework that will allow a group of people access to a database that includes primary texts, comments about them, and previous requests that have been made for information; our initial work has been with a legal text containing rules for arbitration.

Our major focus is on understanding how people working as scientists and professionals on problems in their area of expertise actually use information to solve those problems. The strategy we are pursuing entails constructing computer-based systems in which the users can access a variety of different kinds of information through dialogue interactions in ordinary conversational language. By observing how they are used we expect to be able to guide modifications for making them successively more and more valuable.

The information-retrieval systems of information science and the knowledge-based expert systems of artificial intelligence can be viewed as constituting two ends of a continuum of facilities relevant for knowledge synthesis and interpretation. Considered in idealized form, both represent static states, the content of information-retrieval

systems providing the raw materials from which people derive information relevant for their needs; the expert systems embodying digested knowledge consensually validated as relevant for some area of inquiry. In contrast, the systems we are developing, which might be called "systems for experts," fall somewhere in between these extremes. Accordingly, they reflect the dynamic instabilities and uncertainties of the continuing search for new ideas and new answers, the core of the problems of knowledge synthesis and interpretation.

E. REFERENCES

Allen, J; and Perrault, CR. "Analyzing Intention in Dialogues." Artificial Intelligence 15 (1980): 143-178.

Bates, MJ. "Information Search Tactics." Journal of the American Society for Information Science 30 (1979): 205-214. (a)

Bates, MJ. "Idea Tactics." Journal of the American Society for Information Science 30 (1979): 280-289. (b)

Belkin, NJ. "Information Concepts for Information Science." Journal of Documentation 34 (1978):55-85.

Belkin, NJ. "Anomalous States of Knowledge as a Basis for Information Retrieval." Canadian Journal of Information Science 5 (1980):133-143.

Belkin, NJ; Brooks, HM; and Oddy, RN. "Representation and Classification of Knowledge and Information for Use in Interactive Information Retrieval." In IRFIS 3: Proceedings of the Third International Research Forum in Information Science, pp. 146-185. Oslo, Norway: Norwegian Library School, 1979.

Bernstein, LM; Siegel, ER; and Ford, WH. "The Hepatitis Knowledge Base: A Prototype Information Transfer System." Annals of Internal Medicine 93 (1980):165-222.

Bobrow, D, and Winograd, T. "An Overview of KRL, A Knowledge Representation Language." Cognitive Science 1 (1977):3-46.

Bourne, CP. "On-Line Systems: History, Technology, and Economics." Journal of the American Society for Information Science 31 (1980):155-160.

Brachman, RJ. "On the Epistemological Status of Semantic Networks." In Associative Networks - The Representation and Use of Knowledge in Computers, edited by NV Findler, pp. 3-50. New York: Academic Press, 1979.

Brachman, RJ; and Smith, BC. "Special Issue on Knowledge Representation." ACM SIGART Newsletter No. 70 (1980).

Bronowski, J. Nature of Knowledge: The Philosophy of Contemporary Science. New York: Science Books, 1969.

Brown, JS; and Burton, RR. "Multiple Representations of Knowledge for Tutorial Reasoning." In Representation and Understanding, edited by DG Bobrow and A Collins, pp. 311-349. New York: Academic Press, 1975.

Chandioux, J. "METEO: an Operational System for the Translation of Public Weather Forecasts." American Journal of Computational Linguistics Microfiche 46 (1976):27-36.

Codd, EF. "A Relational Model for Large Shared Data Banks." Communications of the ACM 13 (1970):377-387.

Codd, EF. "Seven Steps to Rendezvous with the Casual User." In Data Base Management, edited by JW Klimbie and KI Koffeman, pp. 179-200. Amsterdam: North-Holland, 1974.

Codd, EF. "How About Recently? (English Dialogue with Relational Databases Using RENDEZVOUS Version 1)." In Databases: Improving Usability and Responsiveness, edited by B Shneiderman, pp. 3-28. New York: Academic Press, 1978.

Cohen, H. "What Is an Image?" Sixth International Joint Conference on Artificial Intelligence, pp. 1028-1057. Stanford University, 1979.

Cohen, P. "On knowing what to say: planning speech acts." Technical Report No. 118, Department of Computer Science, University of Toronto, January 1978.

Cohen, PR, and Perrault, CR. "Elements of a Plan-Based Theory of Speech Acts." Cognitive Science 3 (1979):177-212.

Collins, A. "Why Cognitive Science." Cognitive Science 1 (1977):1-2.

Cooper, WS; and Maron, ME. "Foundations of Probabilistic and Utility-Theoretic Indexing." Journal of the Association for Computing Machinery 25 (1978):67-80.

Crane, D. Invisible Colleges. Chicago: University of Chicago Press, 1972.

Damerau, FJ. "Automated Language Processing." In Annual Review of Information Science and Technology, Volume 11, edited by ME Williams, pp. 107-161. Washington, DC: American Society for Information Science, 1976.

Davis, R. "Interactive Transfer of Expertise: Acquisition of New Inference Rules." Artificial Intelligence 12 (1979):121-157.

Davis, R; and King, J. "An Overview of Production Systems." In Machine Representation of Knowledge, edited by EW Elcock and D Michie, pp. 300-332. New York: Wiley, 1976.

Doszkocs, TE; and Rapp, BA. "Searching MEDLINE in English: A Prototype User Interface with Natural Language Query, Ranked Output, and Relevance Feedback." Proceedings of the ASIS Annual Meeting, Volume 16, edited by RD Tally and RR Deultgen, pp. 131-139. New York: Knowledge Industry Publications, 1979.

Duda, R; Gaschnig, J; and Hart, P. "Model Design in the Prospector Consulting System for Mineral Exploration." In Expert Systems in the Micro-electronic Age, edited by D Michie, pp. 153-167. Edinburgh: Edinburgh University Press, 1979.

Duda, R; Hart, P; and Nilsson, NJ. "Subjective Bayesian Methods for Rule-Based Inference Systems." Proceedings of the National Computer Conference, Volume 45, pp. 1075-1082. Arlington, Virginia: AFIPS Press, 1976.

Epstein, MN. Natural Language Access to Clinical Data Bases. Ph.D. Dissertation, University of California, San Francisco, 1980.

Epstein, MN; and Walker, DE. "Natural Language Access to a Melanoma Data Base." Proceedings of The Second Annual Symposium on Computer Application in Medical Care, pp. 320-325. New York: IEEE, 1978.

Feigenbaum, EA; Buchanan, BG; and Lederberg, J. On Generality and Problem Solving: A Case Study Using the DENDRAL Program. In Machine Intelligence, Volume 6, edited by B Meltzer and D Michie, pp. 165-190. Edinburgh: Edinburgh University Press, 1971.

Feigenbaum, EA. "The Art of Artificial Intelligence--Themes and Case Studies of Knowledge Engineering." Proceedings of the National Computer Conference, Volume 47, pp. 227-240. Montvale, New Jersey: AFIPS Press, 1978. Also published as "Themes and Case Studies of Knowledge Engineering." In Expert Systems in the Micro-electronic Age, edited by D Michie, pp. 3-25. Edinburgh: Edinburgh University Press, 1979.

Fikes, RE; and Hendrix, GG. "A Network-Based Knowledge Representation and its Natural Deduction System." Proceedings of the Fifth International Joint Conference on Artificial Intelligence, pp. 235-246. Carnegie-Mellon University, 1977.

Fillmore, CJ. "The Case for Case." In Universals in Linguistic Theory, edited by E Bach and R Harms, pp 1-90. New York: Holt, Rinehart, and Winston, 1968.

Findler, NV, Ed. Associative Networks - The Representation and Use of Knowledge in Computers. New York: Academic Press, 1979.

Fox, MS; and Palay, AJ. "The BROWSE System: An Introduction." In Information Choices and Policies: Proceedings of the 42nd Annual Meeting of the American Society for Information Science, pp. 183-193. White Plains, New York: Knowledge Industry Publications, 1979.

Gadamer, H-G. Philosophical Hermeneutics. Berkeley: University of California Press, 1976.

Green, CC. The Application of Theorem Proving to Question Answering. Artificial Intelligence Project Memo AI-96, Stanford University, 1969.

Grishman, R; and Hirschman, L. "Question Answering from Natural Language Medical Data Bases." Artificial Intelligence 11 (1978):25-43.

Grosz, BJ. "The Representation and Use of Focus in a System for Understanding Dialogs." Proceedings of the Fifth International Joint Conference on Artificial Intelligence, pp. 67-76. Carnegie-Mellon University, 1977.

Grosz, BJ. "Utterance and Objective: Issues in Natural Language Communication." Sixth International Joint Conference on Artificial Intelligence, pp. 1067-1076. Stanford University, 1979.

Grosz, BJ. "Focusing and Description in Natural Language Dialogues." In Elements of Discourse Understanding: Proceedings of a Workshop on Computational Aspects of Linguistic Structure and Discourse Setting, edited by AK Joshi, I Sag, and BL Webber, pp. 84-105. Cambridge: Cambridge University Press, 1981.

Haas, N; and Hendrix, GG. "An Approach to Acquiring and Applying Knowledge." Proceedings of the First Annual National Conference on Artificial Intelligence, pp. 235-239. Stanford University, 1980.

Hall, JL; and Brown, MJ. Online Bibliographic Databases: An International Directory. Second Edition. London: Aslib, 1981.

Halliday, MA. "Notes on Transitivity and Theme in English. Part 2." Journal of Linguistics 31 (1967):177-274.

Harris, LR. "User Oriented Data Base Query with the ROBOT Natural Language Query System." International Journal of Man Machine Studies 9 (1977):697-713.

Harris, LR. "Experience with ROBOT in 12 Commercial Natural Language Data Base Query Applications." Proceedings of the Sixth International Joint Conference on Artificial Intelligence, pp. 365-368. Stanford University, 1979.

Hart, PE; Duda, RO; and Einaudi MT. "PROSPECTOR--A Computer-Based Consultation System for Mineral Exploration." Journal of the International Association for Mathematical Geology 10 (1978):589-610.

Harter, SP. "Statistical Approaches to Automatic Indexing." Drexel Library Quarterly 14 (1978):57-74.

Hayes, P; and Mouradian, G. "Flexible Parsing." In Proceedings of the 18th Annual Meeting of the Association for Computational Linguistics, pp. 97-103. University of Pennsylvania, 1980.

Hayes, P; and Reddy, R. "Graceful Interaction in Man-Machine Communication." Sixth International Joint Conference on Artificial Intelligence, pp. 372-374. Stanford University, 1979.

Hayes-Roth, F; Waterman, DA; and Lenat, DB. "Principles of Pattern-Directed Inference Systems." In Pattern-Directed Inference Systems, edited by DA Waterman and F Hayes-Roth, pp. 577-601. New York: Academic Press, 1978.

Hendrix, GG. "Human Engineering for Applied Natural Language Processing." Proceedings of the Fifth International Joint Conference on Artificial Intelligence, pp. 183-191. Carnegie-Mellon University, 1977. (a)

Hendrix, GG. The LIFER Manual: A Guide to Building Practical Natural Language Interfaces. Technical Note 138, Artificial Intelligence Center, Stanford Research Institute, Menlo Park, California, February 1977. (b)

Hendrix, GG. "Encoding Knowledge in Partitioned Networks." In Associative Networks - The Representation and Use of Knowledge in Computers, edited by NV Findler, pp. 51-92. New York: Academic Press, 1979.

Hendrix, GG; Sacerdoti, ED; Sagalowicz, D; and Slocum, J. "Developing a Natural Language Interface to Complex Data." ACM Transactions on Database Systems 3 (1978):105-147.

Hobbs, J. "Coherence and Interpretation in English Texts." Proceedings of the Fifth International Joint Conference on Artificial Intelligence, pp. 110-116. Carnegie-Mellon University, 1977. (a)

Hobbs, J. "From English Descriptions of Algorithms into Programs." Proceedings, Annual Conference, Association for Computing Machinery, pp. 323-329. New York: ACM, 1977. (b)

Hutchins, WJ. "Machine Translation and Machine-Aided Translation." Journal of Documentation 34 (1978):119-159.

Jespersen, O. A Modern English Grammar on Historical Principles. London: Allen and Unwin, 1914-1929.

Kaplan, SJ. Cooperative Responses from a Portable Natural Language Data Base Query System. Ph.D. Dissertation, University of Pennsylvania, 1979.

Karlgren, H; and Walker, DE. "The POLYTEXT System - A New Design for a Text Retrieval System." To be published in the Proceedings of a Conference on Questions and Answers held in Visegrad, Hungary, 4-6 May 1980.

Karlgren, H; Walker, DE; and Kay, M. Computer Aids in Translation. In preparation.

Kiefer, F. Topic-Comment Structure of Texts: Some Preliminary Remarks. Report, Kval Institute for Information Science, Stockholm, September 1979.

Kochen, M. Principles of Information Retrieval. Los Angeles: Melville, 1974.

Kuhn, T. The Structure of Scientific Revolutions. Second Edition, Enlarged. Chicago: University of Chicago Press, 1970.

Kwasny, S; and Sondheimer, NK. "Relaxation Techniques for Parsing Grammatically Ill-Formed Input in Natural Language Understanding Systems." American Journal of Computational Linguistics 7 (1981):99-108.

Lancaster, FW. Toward Paperless Information Systems. New York: Academic Press, 1978.

Lancaster, FW. Information Retrieval Systems. Second Edition. New York: Wiley, 1979.

Lancaster, FW; and Fayen, EG. Information Retrieval On-Line. Los Angeles: Melville, 1973.

Landau, RN; Wanger, J; and Berger, MC. Directory of Online Databases, Volume 1, Number 1. Santa Monica: Cuadra Associates, 1979.

Landsbergen, SPJ. "Syntax and Formal Semantics of English in PHLIQA1." In Coling 76, Preprints of the 6th International Conference on Computational Linguistics, No. 21. Ottawa, 1976.

Larson, S; and Williams, ME. "Computer Assisted Legal Research." In Annual Review of Information Science and Technology, Volume 15, edited by ME Williams, pp. 251-286. White Plains, New York: Knowledge Industry Publications, 1980.

Lea, W. Trends in Speech Recognition. Englewood Cliffs, New Jersey: Prentice-Hall, 1980.

Loef, S. The POLYTEXT/ARBIT Demonstration System. FOA Report C40121-M7, Swedish National Defence Research Institute, Umea, Sweden, September 1980.

Loveland, DW. Automated Theorem Proving: A Logical Basis. New York: North Holland, 1978.

Martin, TH. A Feature Analysis of Interactive Retrieval Systems. Institute for Communications Research, Stanford University, 1974.

Mays, E. "Failures in Natural Language Systems: Applications to Data Base Query Systems." In Proceedings of the First Annual National Conference on Artificial Intelligence, pp. 317-330. Stanford University, 1980.

McCarthy, J; and Hayes, PJ. "Some Philosophical Problems from the Standpoint of Artificial Intelligence." In Machine Intelligence, Volume 4, edited by B Meltzer and D Michie, pp. 463-502. Edinburgh: Edinburgh University Press, 1969.

McDermott, D; and Doyle, J. "Non-Monotonic Logic I." Artificial Intelligence 13 (1980):41-72.

McGill, MJ; and Huitfeldt, J. "Experimental Techniques of Information Retrieval." In Annual Review of Information Science and Technology, Volume 14, edited by ME Williams, pp. 93-127. New York: Knowledge Industry Publications, 1979.

McKeown, KR. "Generating Relevant Explanations: Natural Language Responses to Questions about Database Structures." In Proceedings of the First Annual National Conference on Artificial Intelligence, pp. 306-309. Stanford University, 1980.

Minker, J. Information Storage and Retrieval: A Survey and Functional Description. ACM SIGIR Forum 12 (1977):1-108.

Minsky, M. A Framework for Representing Knowledge. In The Psychology of Computer Vision, edited by P Winston, pp. 211-280. New York: McGraw-Hill, 1975.

Moore, R. Reasoning about Knowledge and Actions. Technical Note 191, Artificial Intelligence Center, SRI International, Menlo Park, Calfifornia, 1980.

Newell, A; and Simon, HA. Human Problem Solving. Englewood Cliffs, New Jersey: Prentice-Hall, 1972.

Nii, HP; and Aiello, N. "AGE (Attempt to Generalize): A Knowledge-Based Program for Building Knowledge-Based Programs." Proceedings of the Sixth International Joint Conference on Artificial Intelligence, pp. 645-655. Stanford University, 1979.

Nilsson, NJ. Principles of Artificial Intelligence. Palo Alto, California: Tioga Press, 1980.

O'Connor, J. "Answer Passage Retrieval by Text Searching." Journal of the American Society for Information Science 31 (1980):227-239.

Oddy, RN. "Information Retrieval through Man-Machine Dialogue." Journal of Documentation 33 (1977):1-44.

Palmer, RE. Hermeneutics: Interpretation Theory in Schleiermacher, Dilthey, Heidegger, and Gadamer. Evanston: Northwestern University Press, 1969.

Petrick, SR. "Automatic Syntactic and Semantic Analysis." In Proceedings of the Interdisciplinary Conference on Automated Text Processing, (Bielefeld, German Federal Republic, 8-12 November 1978), edited by J Petofi and S Allen. Dordrecht, Holland: Reidel, in press.

Pople, H. "The Formation of Composite Hypotheses in Diagnostic Problem Solving: An Exercise in Synthetic Reasoning." Proceedings of the Fifth International Joint Conference on Artificial Intelligence, pp. 1030-1037. Carnegie-Mellon University, 1977.

Pople, H; Myers, JD; and Miller, RA. "DIALOG: A Model of Diagnostic Logic for Internal Medicine." Proceedings of the Fourth International Joint Conference on Artificial Intelligence, pp. 848-855. Massachusetts Institute of Technology, 1975.

Price, DJdeS. Science Since Babylon. New Haven: Yale University Press, 1968.

Quillian, MR. "Semantic Memory." In Semantic Information Processing, edited by M Minsky, pp. 227-270. Cambridge, Massachusetts: Press, 1968.

Reboh, R. Knowledge-Engineering Techniques and Tools in the Prospector Environment. Technical Note 243, Artificial Intelligence Center, SRI International, Menlo Park, California, 1981.

Reddy, MJ. "The Conduit Metaphor--A Case of Frame Conflict in Our Language about Language." In Metaphor and Thought, edited by A Artony, pp. 284-324. Cambridge: Cambridge University Press, 1979.

Robertson, G; McCracken, D; and Newell, A. The ZOG Approach to Man-Machine Communication. CMU-CS-79-148, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Pennsylvania, October 1979.

Robinson, AE; Appelt, DE; Grosz, BJ; Hendrix, GG; and Robinson, JJ. "Interpreting Natural-Language Utterances in Dialog About Tasks." Communications of the ACM, in press. SRI Technical Note 210, Artificial Intelligence Center, SRI International, Menlo Park, California. March 1980.

Robinson, JA. Logic: Form and Function. New York: North Holland, 1979.

Robinson, JJ. "DIAGRAM: A Grammar for Dialogues." Communications of the ACM, in press. SRI Technical Note 205, Artificial Intelligence Center, SRI International, Menlo Park, California, February 1980.

Rubin, AD. "A Theoretical Taxonomy of the Differences Between Oral and Written Language." In Theoretical Issues in Reading Comprehension, edited by R Spiro, B Bruce, and W Brewer. Hillsdale, New York: Lawrence Erlbaum, 1978.

Sager, N. Natural Language Information Processing: A Computer Grammar of English and Its Applications. Reading, Massachusetts: Addison-Wesley, 1981.

Schank, R. "Interestingness: Controlling Inferences." Artificial Intelligence 12 (1979):273-297.

Schank, R; and Abelson, R. Scripts, Plans, Goals, and Understanding. Hillsdale, New Jersey: Lawrence Erlbaum, 1977.

Schank, R; Lebowitz, M; and Birnbaum, L. "An Integrated Understander." American Journal of Computational Linguistics 6 (1980):13-30.

Schank, R; and Nash-Webber, BL. (Eds.) Theoretical Issues in Natural Language Processing. Massachusetts Institute of Technology, 1975.

Schoen, DA. "Generative Metaphor: A Perspective on Problem-Setting in Social Policy." In Metaphor and Thought, edited by A Artony, pp. 254-283. Cambridge: Cambridge University Press, 1979.

Sgall, P. "Perspective Paper: Linguistics." In Natural Language in Information Science, edited by DE Walker, H Karlgren, and M Kay, pp. 101-126. Stockholm: Skriptor, 1977.

Shortliffe, EH. Computer-Based Medical Consultations: MYCIN. New York: Elsevier/North-Holland, 1976.

Shortliffe, EH; Buchanan, BG; and Feigenbaum, EA. "Knowledge Engineering for Medical Decision Making." Proceedings of the IEEE 67 (1967):1207-1224.

Sidner, CL. Toward a Computational Theory of Definite Anaphora Comprehension in English Discourse. Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1979.

Simmons, RA. "Natural Language Question-Answering Systems: 1969." Communications of the ACM 13 (1970):15-30.

Simmons, RA. "Semantic Networks: Their Computation and Use for Understanding English." In Computer Models of Thought and Language, edited by RC Schank and KM Colby, pp. 63-113. San Francisco: Freeman, 1973.

Smith, LC. "Artificial Intelligence Applications in Information Systems." In Annual Review of Information Science and Technology, Volume 15, edited by ME Williams, pp. 67-105. White Plains, New York: Knowledge Industry Publications, 1980.

Sparck Jones, K; and Bates, RG. Research on Automatic Indexing 1974-1976. Computer Laboratory, University of Cambridge, Cambridge, England, 1977.

Sparck Jones, K; and Kay, M. Linguistics and Information Science. New York: Academic Press, 1973.

Sprowl, J. "Computer-Assisted Legal Research--An Analysis of Full-Text Document Retrieval Systems, Particularly the LEXIS System." American Bar Foundation Research Journal 175 (1976):175-226.

Stockholm Chamber of Commerce. Arbitration in Sweden. Stockholm: Wetter and Wetter, 1977.

Templeton, M. "EUFID: A Friendly and Flexible Front-end for Data Management Systems." Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics, pp. 91-93. University of California at San Diego, 11-12 August, 1979.

Thompson, BH; and Thompson, FB. "Rapidly Extendable Natural Language." Proceedings of the 1978 Annual Meeting of the Association for Computing Machinery, pp. 173-182. New York: ACM, 1978.

Thompson, FB; and Thompson, BH. "Practical Natural Language Processing: The REL System as Prototype." In Advances in Computers, Volume 13, edited by M Rubinoff and MC Yovits, pp. 109-168. New York: Academic Press, 1975.

van Rijsbergen, CJ. Information Retrieval. Second edition. London: Butterworths, 1979.

Vickery, BC. On Retrieval System Theory. London: Butterworths, 1961.

Walker, DE. "Automated Language Processing." In Annual Review of Information Science and Technology, Volume 8, edited by CA Cuadra and AW Luke, pp. 69-119. Washington, DC: American Society for Information Science, 1973.

Walker, DE, Ed. Understanding Spoken Language. New York: North-Holland, 1978.

Walker, DE; and Hobbs, JR. "Natural Language Access to Medical Text." Proceedings of the Fifth Annual Symposium on Computer Applications in Medical Care. IEEE, New York, 1981.

Waltz, DL. An English Language Question Answering System for a Large Relational Data Base. Communications of the ACM 21 (1978):526-539. (a)

Waltz, DL. (Ed.) TINLAP-2: Theoretical Issues in Natural Language Processing-2. University of Illinois, Urbana-Champaign, 1978. (b)

Wanger, J; and Landau, RN. "Nonbibliographic On-Line Data Base Services." Journal of the American Society for Information Science 31 (1980):171-180.

Waterman, DA; and Hayes-Roth, F. Pattern-Directed Inference Systems. New York: Academic Press, 1978.

Watt, WC. "Habitability." American Documentation 19 (1968):338-351.

Webber, BL. A Formal Approach to Discourse Anaphora. TN-3761, Bolt Beranek and Newman, Cambridge, Massachusetts, 1978.

Weischedel, RM; and Black, JE. "Responding Intelligently to Unparsable Inputs." American Journal of Computational Linguistics 6 (1980):97-109.

Weiss, SM; Kulikowski, CA; Amarel, S; and Safir, A. "A Model-Based Method for Computer-Aided Medical Decision Making." Artificial Intelligence 11 (1978):145-172.

Williams, ME (Ed). Computer-Readable Data Bases: A Directory and Data Source Book. White Plains, New York: Knowledge Industry Publications, 1979.

Williams, ME (Ed). Annual Review of Information Science and Technology, Volume 15. White Plains, New York: Knowledge Industry Publications, 1980.

Winograd, T. "Frame Representations and the Procedural-Declarative Controversy." In Representation and Understanding, edited by DG Bobrow and A Collins, pp. 185-210. New York: Academic Press, 1975.

Winograd, T. "What Does It Mean to Understand Language?" Cognitive Science 4 (1980):209-241.

Winston, PH. Artificial Intelligence. Reading, Massachusetts: Addison-Wesley, 1977.

Woods, WA. "What's In a Link? Foundations for Semantic Networks." In Representation and Understanding, edited by DG Bobrow and A Collins, pp. 35-82. New York: Academic Press, 1975.

Woods, WA; Kaplan, RM; and Nash-Webber, B. The Lunar Sciences Natural Language Information System. BBN Report 2378, Bolt Beranek and Newman, Cambridge, Massachusetts, 1972.

Zachary, WW. "A Survey of Approaches and Issues in Machine-Aided Translation Systems." Computers and the Humanities 13 (1979):17-28.