Final Report on the Automated Classification and Retrieval Project (MedSORT-I)

Jaime G. Carbonell, David A. Evans,
Dana S. Scott, and Richmond H. Thomason

December 1985

Report No. CMU-LCL-85-1

Laboratory for Computational Linguistics

139 Baker Hall
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213

Final Report on the
Automated Classification and Retrieval Project

Grant: N01-LM-4-3529

National Library of Medicine
Bethesda, Maryland

The MedSORT Project

Principal Investigators:

Jaime G. Carbonell
David A. Evans
Dana S. Scott
Richmond H. Thomason

Departments of Philosophy and Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania

December 1985

Final Report on the Automated Classification and Retrieval Project

The MedSORT Project
Carnegie-Mellon University

Executive Summary. The work of this project was supported on a one-year non-renewable contract (Grant: N01-LM-4-3529) funded for a period running from 28 September 1984 through 27 September 1985. The contractors were additionally granted an extension of time in submitting the written final report to 1 December 1985. The report has two main divisions: (1) an essay in five parts on *project background, methodology, conclusions and recommendations, project activities*, and a *guide to the appendices*, and (2) 2400 pages of *supporting documentation* of the work carried out on the project, presented in a comprehensive array of technical appendices.

The discussion of project background (Part I) reviews the conditions of the original Request for Proposal (RFP No. NLM-84-115/PSP) and the selection of the subject domain *(The Management of Rheumatoid Disorders)*. There is also a section of personal and institutional acknowledgments. Part II on methodology begins with a statement about *organizing biomedical information* and the distinctions being drawn between the concepts: *lexicon, thesuarus, dictionary, knowledge base, expert system,* and *information base.* A section on the relation between *knowledge representation* and *indexing systems* is followed by a technical discussions of the approach to knowledge representation used for the project and of the methods of *natural-language processing* being employed. The last section concerns the extensive project *thesaurus work.*

Part HI presents the conclusions and recommendations reached in the course of the project. The first section has a very **full** statement on the *need* for natural-language processing and on the *problems* of domain coverage. The second section outlines three *objectives* that are proposed for the continuation of this project: (1) *standardizing* biomedical knowledge bases, (2) developing a *medical indexer's workbench,* and (3) making progress on *basic linguistic research.* The last two sections summarize in a synoptic form the *conclusions* and *recommendations.* The three most important conclusions are: (a) one of the most natural applications of combining detailed knowledge representation with natural-language processing is in the *semi-automated indexing* of free text; (b) long-range success in projects in automated, intelligent medical indexing is heavily dependent on a research environment in which useful *tools* can be developed and tested; (c) cooperation with *medical experts* is essential both for research and development in biomedical informatics. In fact, the principal investigators intend to continue collaborating with more than one group of medical experts. A very long-term objective is to build *high quality, extensible expert systems* on top of the developing knowledge base. The continuation of the MedSORT Project is an essential step in that direction.

Table of Contents

Final Report on the Automated Classification and Retrieval Project

The MedSORT Project

Carnegie-Mellon University

This report has two main divisions: (1) an essay in five parts on project background, methodology, conclusions and recommendations, project activities, and a guide to the appendices (this document), and (2) 2400 pages of supporting documentation of the work carried out on the project, presented in a comprehensive array of technical appendices.

Part I. Project Background

The work of this project was supported on a one-year non-renewable contract funded for a period running from 28 September 1984 through 27 September 1985. The contractors were additionally granted an extension of time in submitting the written final report to 1 December 1985.

The *deliverables* of the contract were to be two oral presentations (and subsequent summaries) following each six-month period and a written final report. The stated purpose of the final report was to provide a summation of the work performed and results obtained for the entire contract period of performance. According to the contract, the report has to be written in sufficient detail to describe comprehensively the results achieved, including a complete description of the underlying methodology (algorithms, theory, references, and other supporting information) as well as conclusions and recommendations.

The original Request for Proposal (RFP No. NLM-84-115/PSP) from the National Library of Medicine was dated 29 June 1984. The proposal from the Principal Investigators was submitted at the end of July 1984 and was entitled *Toward the Automation of Content-Access Methods for Large-Scale Textual Databases.* This proposal was subsequently accepted after a presentation in person by the Principal Investigators at the Lister Hill National Center for Biomedical Communications. After the grant was awarded, the contractors applied for and obtained permission to refer to their project more briefly as **The MedSORT Project**. Here **SORT** refers to *Subject-Oriented Retrieval of Text*, and **Med** to the (bio)medical domain. The Principal Investigators are continuing related studies beyond the original contract period and reserve the project name to cover these activities at Carnegie-Mellon University.

1. Objectives of the Project

The *Statement of Work* accompanying the original RFP required the contractor to conduct studies, analyses, and evaluations in selected areas of computer science directed at solving problems related to automated text processing, knowledge classification, knowledge representation, and retrieval of published biomedical literature. Within this context, two general areas of research were laid out:

1. Investigation of knowledge representation and natural-language understanding methods.
2. Identification, development, and evaluation of software techniques and languages to both support the research and to enable its application to the construction of an automated system.

The contractors believe that they have made essential contributions to both of these general areas; these contributions are explained in detail in subsequent sections of this report.

The *Statement of Work* further specified that the scope of the research may range from basic or theoretical research in the above areas to applied research and development of prototype systems and methods. These areas were further elaborated in the *Statement,* and it was emphasized that they be considered in an integrated perspective within the contexts of the goal of the project. Specifically desired were the following:

1. *Development of a syntactico-semantic methodology capable of producing a formal semantic representation of the texts.* That formal representation will be stored in the information database. Such a natural-language understanding method also involves the following interdependent processes: *lexical ellipses and morpho-lcxical analysis; resolution of paraphrases, ambiguities, lexical ellipses and anaphora; recovering of implicit information; and discourse analysis.*
2. *Creation, augmentation, and updating of a knowledge base.* The research shall focus on appropriate methods for knowledge representation such as, but not necessarily limited to, *frame-based representation, semantic networks, rule-based systems, and predicate logic.* Problems to be considered include: *how the various concepts and relationships that constitute the domain of the literature can be formally denoted; what models, formalisms, or meta-theory can be constructed to describe the knowledge representation process; and what can be used to obtain measures of the effectiveness of such representations.*

3. *Creation, augmentation, and interrogation of the information database for the purpose of retrieving semantically coded information.* At present, the knowledge base and the information database are separate concepts. The former is to be considered as a validated set of generic knowledge or theory, while the latter is to be considered as a collection of facts.

The contractors feel fully justified in claiming that they have pursued these project aims to the letter; all these areas of investigation—some necessarily pursued more vigorously than others—will be commented on within this report.

Needless to say, the topics laid out cover *all* the basic problems in the fields of natural-language processing, indexing, and retrieval. On the side of the contractors it has been somewhat difficult to assess exactly what amount of progress the sponsor of this project actually expected over a brief, one-year period. The present report therefore explains what the MedSORT Group has done in the period, and the contractors would hope that some explicit, detailed, written evaluation of progress will be communicated to them from the Lister Hill National Center or their advisors.

The *Statement of Work* additionally included the comment:

> A possible area of research, for the future, is to address the connection between the information database and the knowledge base with respect to consistency between the two, validation of facts, and augmentation and updating of the knowledge base.

The contractors certainly agree that this is a vital issue, and indeed they are endeavoring to create in their own Pittsburgh research environment the necessary connections between medical experts (in selected subject domains) and computer scientists that would help make the essential activity of *validation* possible. To attempt to do this for all of biomedical knowledge, however, is the work of many, many lifetimes! It seems most reasonable to envision this kind of work being carried out under the general direction and coordination of the National Library of Medicine, but it is obviously a major task to set this in motion involving many agencies and bodies, both professional and political. The MedSORT Group has further comments and some related, shorter-term, smaller-scale recommendations to make; these have been included in the Part III of this report.

2. Selection of a Subject Domain

The *Statement of Work* further stated that the Lister Hill National Center for Biomedical Communications was to specify or approve all subject domains used to develop knowledge structures. At the time of writing the RFP it was thought that several grants would be awarded, and the requirement of having a focal-point rôle was included explicitly for the purpose of providing compatibility between independent research projects. In selecting the subject domain, the following guidelines were to be observed:.

1. The domain should relate to the clinical practice of medicine in the United States.
2. The domain should be of immediate and ongoing interest to the medical community of the United States.
3. Information about the domain should be published in high quality journals that are readily available to the American practitioner.
4. Consideration should be given to the complexity of the knowledge base that would be required by the domain.

The subject domain finally chosen was **The Management of Rheumatoid Disorders.** The contractors regard this choice a most interesting one, and they have learned a very great amount about construction of knowledge bases from the ensuing investigation of the 'problems of this area. Moreover, this proved to be a rich and challenging subject domain, enabling us to make far more progress in developing knowledge representation tools than might otherwise have been possible in a one-year project. With hindsight we can now see how the very considerable breadth of the domain directed our research efforts into all the core aspects of medical-knowledge organization. In one year, however, only the areas most pertinent to rheumatoid arthritis could be treated in substantial depth.

3. Resources and Acknowledgements

The contractors would like to record their thanks both to Dr. Donald A. B. Lindberg, M.D., Director National Library of Medicine and to Dr. Harold M. Schoolman, M.D., Deputy Director for Research and Education, for their personal interest in this project and for the advice they offered during the contract period. Thanks axe owing, too, to Earl B. Henderson, Acting Director, Lister Hill National Center, who acted as Project Officer for this contract.

Special thanks go to Susanne Humphrey, Lister Hill National Center, without whose consultation and assistance the project would not have been possible. Her own contribution is recorded in Appendix B.4, but her help in understanding MeSH and many issues of indexing practice extends throughout all the work of the project. We would also like to thank Dr. Madeleine Graitson for her interest and assistance in obtaining resources, especially SNOMED, and Jules P. Aronson for his help in setting up computer communications and obtaining certain MeSH files for us. We are particularly grateful to Ms. Beth Van Lenten, Acting Head of the Indexing Section at the National Library of Medicine, for her generous cooperation and aid in arranging for our protocol studies of National Library of Medicine indexing practice (reported in Appendix A.4).

In particular, the contractors wish to record here their very warm appreciation to Dr. Thomas A. Medsger, Jr., M.D., Professor of Medicine, Rheumatology, and Clinical Immunology at the University of Pittsburgh. Dr. Medsger gave unselfishly of his time and advice over the whole contract period, and he was particularly patient and helpful in discussing and reviewing the work of the project staff (who were trained in artificial intelligence, linguistics, and computational linguistics, rather than medicine). Dr. Medsger's commitment and interest to this work was essential and particularly encouraging, and the MedSORT group hopes the spirit of cooperation forged will continue in work on other, related projects. In addition we would also like to thank Dr. Rosalind Ramsey-Goldman, M.D., for her assistance while Dr. Medsger was away.

We have received much help and encouragement from tl\e members of the-INTERNIST Project and the Decision Systems Laboratory at the University of Pittsburgh School of Medicine. Our acquisition of the INTERNIST-1 knowledge base, incorporated in our MedSORT thesaurus, was made possible through the generosity of Drs. Randolph Miller, M.D., Jack D. Myers, M.D., and Fred Masarie, M.D., who cooperated with us in a spirit of friendship and in the interest of furthering research in biomedical informatics. Many times, as well, our project benefited from discussions with Drs. Harry E. Pople, Jr., John K. Vries, M.D., and Gordon E. Banks, M.D. We have been saved from many pitfalls by their frequent good counsel.

The Department of Computer Science at Carnegie-Mellon University has contributed many resources to our Project beyond those initially projected in our project budget. We would like to thank the Department for its extensive support and would especially like to thank Allen Stoltzfus for helping us keep track of numerous project details. Secretarial assistance for the Project was very ably provided by Mrs. Lydia Defilippo.

Finally, for their many efforts and tireless dedication to our work, we would like to thank the graduate students, research assistants, and programmers who comprised the MedSORT Project Team at various times: Armar A. Archbold, John Aronis, Andi Blaustein, Renee Bornstein, Carol J. Crain, Charles Cross, Robert Frederking, Debra E. Graham, John F. Horty, Robert Joseph, Sandra Katz, Marion Kee, Benjamin MacLaren, Ira Monarch, Stephen E. Morrisson, Shinako Ogawara, and Staci Quackenbush.

Part II. Methodology

«

Necessarily, our first year of MedSORT Project activity has involved much basic research. We have concentrated our efforts on establishing the feasibility of our approach to text processing in the context of indexing and retrieval problems of the sort in evidence at the National Library of Medicine. Consequently, we have been especially concerned with developing a proper understanding of these problems; and with developing the essential tools for an automatic text-processing system in the biomedical domain. For us, these tools have taken the form of a *coherent knowledge representation schema,* an *efficient natural-language parser,* and a *domain-specific thesaurus.* We include here in this part of the report sections on indexing, knowledge representation, natural-language processing, and the MedSORT thesaurus that take up each of these points.

We begin our presentation of methodology with a brief discussion of the distinctions we find essential in the organization of information in the biomedical domain. We then organize our further discussion of MedSORT Project methodology in the following sections around the key concepts—at a fairly 'high' level of description—that distinguish our approach to the problem of providing automatic access to free-text in large-scale databases. In each section we include numerous references to appendices, where the details of our work are to be found. Some of the appendices are in fact additional, independent reports or documents that were prepared in the course of the project. Most, however, contain data and analyses that reflect the day-to-day activity of various MedSORT working groups on indexing, knowledge representation, thesaurus construction, and parser development. Each of these (often large) appendices contains its own brief report and references to subsections.

1. On the Organization of Biomedical Information

We claim that future access to information in the biomedical sciences—at all levels—will depend on the intelligent automation of our information resources; this, in turn, will depend on our ability to *standardize* the ways in which biomedical information is expressed, encoded, and communicated.

In considering these issues, we believe that it is vital to bear in mind the distinction between *natural language* and *knowledge representation.* The latter is encoded computationally, and in many applications will be completely hidden from non-expert users; it is a fundamental design criterion that the knowledge representation should be unambiguous and standardized. Natural language, on the other hand, is a social construct, and as such cannot easily be changed or standardized without providing strong inducements to the language-using community.

The channel controlling standardization is the *natural-language interface* to a computer system. By providing automatically for such trivial variation as that between, say, "pancreatic cancer" and "cancer of the pancreas", a natural-language interface can make the standardization of the internal representation language much easier to achieve. Also by formulating messages in a single way, such interfaces can provide a model of uniformity of terminology, which users can then learn to emulate in other situations.

Natural-language processing not only provides automated translations from natural-language input to canonical machine-held knowledge representations, but also from the representations to natural-language output. *Interpretation* is in general many-one, since natural language provides many ways to say the same thing. *Generation* is usually one-one, since it is typically not necessary for the machine to express the same message in different ways; one clear way to express a concept suffices. This is useful for many reasons, as we have indicated. A computational system, however, is only as useful as the *reasoning* that it can perform; and reasoning cannot take place without a representation that allows information to be encoded unambigously and in enough detail to enable relevant distinctions to be made. Elaborate systems of knowledge representation are often so complex and difficult to use that they are accessible only to highly trained experts. This creates expensive delays in communicating with such systems, makes it difficult to use them interactively, and introduces a source of hard-to-detect errors. The problem is to find the middle way that can be made accessible to users with only a moderate amount of training.

A solution to this problem—and one that has proved useful in many domains—is to work toward the automation of the encoding of large volumes of natural-language input. Natural-language interfaces are especially important in this regard, but they have to be tested in use over extended periods. We contend, moreover, that they are needed for any task requiring expert, controlled descriptions of entities. A system that is able to make more distinctions will be in a better position to perform sophisticated retrieval tasks; and the goal of representing a large sample of natural-language input will ensure that the system has this ability.

In considering the systematic organization of information in any field, it is necessary to keep in mind the distinctions among the following constructs:

- **Lexicon**
- **Thesaurus**
- **Dictionary**
- **Knowledge Base**
- Expert System

A *lexicon* lists the terms that are encountered in the domain in focus, along with information about their syntactic properties (such as *part-of-speech)* and morphology (giving variations such as "x-ray (singulax)/x-rays (plural)/x-radiologic (adjective)/..."). The terms are typically single words, but can include word-fragments (such as the bound morpheme "osteo-") and phrases that are to be considered atomic (such as "Wilson's disease") Naturally, any *practical* lexicon for even highly restricted domains will also embed some portion of a general natural-language lexicon.

The *thesaurus* provides a standard internal terminology in the sense that it establishes the set of relations that are used to distinguish terms, and thus organizes the terms into a *semantic network.* One of the most important reasons for building networks is that properties of terms can be inherited along paths in the network. This automatic accumulation of features is what makes it possible to construct sufficiently detailed terminology lists, where otherwise a complete listing of necessary properties would be infeasible. The thesaurus cannot exist wholly independently of the lexicon, and in most realizations it will repeat some of the information on the grammatical categories of terms. The thesaurus and lexicon together provide the information that is essential to automatic natural-language processing.

A *dictionary,* on the other hand, presents the definitions of terms, with special attention to actual *usage.* It is essential to the human understanding of specialized words. The biomedical field, which is expanding daily, needs a central repository of definitions for facilitating communication at all levels. A dictionary is also needed to help shape the growth of the thesaurus—since none of these information sources is static—and various authorities have to make decisions about the admission of new terms. The pressure to get involved in dictionary building is very strong. In terms of traditional linguistic divisions, then, the essential differences among lexicon, thesaurus, and dictionary are that the lexicon provides *syntactic* information; the thesaurus, *semantic* information; and the dictionary, *pragmatic* information.

A *knowledge base* by contrast is usually limited in scope, aiming at depth of information rather than comprehensiveness. It will go far beyond definitions in becoming a tome in an *encyclopedia,* which would be made up of a suite of knowledge bases. As it gradually grows beyond limited-scope applications, a knowledge base could become a rich, machine-usable, multi-purpose medical information source. Of course, the feasibility of constructing a knowledge base will depend very much on available resources and urgency of applications. The lexicon, thesaurus, and the dictionary, however, axe so basic to any system building involving even superficial natural-language processing, that our claim is that their systematic construction must be undertaken as soon as possible in order to see any progress at all.

Turning to *expert systems*, we claim that these are best built on top of knowledge bases, and that—for the most part—the knowledge base should remain in structure relatively neutral as to the advice the expert system is supposed to produce. That is to say, the expert system should be engaged in *interpreting* the knowledge for purposes of action and decision. It was not the intention of the MedSORT project to engage in the building of such systems; rather we prefer in the first instance to develop *tools* at the more fundamental level of knowledge representation that can help both subsequently to build expert systems on a larger scale than is now possible and to establish a basis for the amalgamation and cooperation of expert systems. We feel in particular that an expert-system builder should be more concerned with the *strategies* of using knowledge than with the *accumulation* of knowledge. The efficient development of useful strategies needs the natural-language tools, however, as that is the only way to make the expert advice humanly understandable.

All of the distinctions we have made—among lexicon, thesaurus, dictionary, knowledge base, and expert system—apply to the 'top' level of any knowledge-based system. Below these, and accesed by them, are *information bases* (typically specialized databases), such as the bibliographic database maintained by the National Library of Medicine. Information bases can have widely varying structure, from simple relational databases to free text to numeric matrices representing imagistic information in the form of pixels. In the design of any natural-language interface to such information bases, both the semantics of the intended representation and the inherent structure and constraints of the information base must be borne in mind. When the information base is textual, the interface problem takes on the characteristics of unrestricted language processing. In the face of such a task, our goal must be to find increasingly complete intermediate solutions in practice, while we develop the top-level knowledge that is prerequisite to full text understanding.

2. Indexing

The design of an indexing system over any database must take account of the knowledge that is represented in the database and the goals and practices of its users—both indexers and searchers. Our first year's effort on establishing the criteria for a partially automated indexing system, therefore, has involved numerous research activities designed to sharpen our understanding of our chosen biomedical knowledge base and the perspectives of its users. Much of our work in this area has influenced our work on knowledge representation, parser design, and thesaurus construction.

Bibliographic indexing and retrieval systems are faced with special problems since the objects in a bibliographic database are potentially extremely complex: free-text titles, abstracts, and, possibly, full research papers. Without a system to process free text, it would be impossible to index all of the information in such a database. The goal must be to index only the subset judged to be of interest or relevance to a user: typically, the principal and key concepts, findings, or results reported in the work being entered. In large databases, even the number of potentially relevant key concepts may be enormous.

The use of any 'language' (e.g., standardized descriptors) over a set of concepts greater than itself will always threaten the quality of an indexing/retrieval system, since quality depends on precision in the use of terminology and on breadth of coverage. Quality is a function of both *consistency in the assignment of descriptors* to items in the database as well as *the discrimination potential* of the descriptors themselves. Paradoxically, however, improvement in one of these parameters can often lead to degradation in the other. Consider just the case of a system employing Single-term[5] descriptors. When the number of possible descriptors is limited, indexer consistency may be high (as there are fewer competing, plausible choices for an indexer to make), but the effect on retrieval will be to return too much irrelevant information. Having too few ways to classify items is certain to lead to the classification of dissimilar items under the same descriptor. When the number of descriptors grows, consistency may suffer (as the number of plausible choices also grows), which in turn can affect retrieval by returning too little relevant information. If there are competing, equally precise terms that contrast only slightly with the specified term, there is no guarantee that all relevant information will be stored under the chosen descriptor.

The best current indexing/retrieval systems do not rely on any strategy as simple as single-term descriptors, of course. Many combine several modes of classifying and accessing information, such as using a thesaurus to control for alternative expressions along with searching on Boolean combinations of keywords. In our investigation of current indexing and retrieval practice, we reviewed the MEDLARS, PRECIS, and SMART systems (cf. Appendix A.I), and reached the following conclusions:

- The use of a thesaurus in structuring descriptor terminology is an essential means of insuring control in consistency and precision. The thesaurus can act as a filter on non-standard expressions and can remind users of allowed alternatives.
- The use of keyword combinations, while insuring breadth of coverage, does not guarantee effective retrieval, especially in large-scale databases. There is no way for string-searching to identify the compatibility of alternative expressions of the same concept, *e.g., increased discomfort after eating* and *painexacerbatedwithfood.*
- The use of descriptors in combinations reflecting linguistic relations provides an efficient and natural means of expressing relevant concepts.

We have argued that the next generation of indexing/retrieval systems must have the ability to represent key concepts in semantically complex but linguistically natural relationships, and must be semi-automated to insure consistency and precision. We believe that a realistic, next-generation system will be one in which there is a division of labor—reflecting a division of expertise—between the user and the machine. The indexer must be free to do what he or she can best do: read and analyze a bibliographic item, and generate a sequence of natural-language phrases that express the key concepts.

The system in turn must automatically process the natural-language expressions, utilizing at least an extensive thesaurus to ensure canonicality. The system-internal indices will be held as case-frames (semantically complex objects), in which the linguistic relations of the orignal expression are preserved, but augmented by additional, domain-specific relationships that are encoded in the thesaurus. In retrieval, the searcher will also be free to express the 'topic* or object of his or her search as a series of natural-language phrases. Using essentially the same mechanism as in indexing, the system will automatically convert the natural-language expressions into case-frames, to be matched against existing indices (in the same form)-

We believe that such a system can be created with existing artificial intelligence technology. In this year of the project, we have created many of the tools that would have to be incorporated into such a system, and we have demonstrated the feasibility of the overall approach by creating a prototype capable of parsing titles. The details will be found in Part II Section 3 and Part II Section 4.

We have naturally been concerned in our first year of work with testing the feasibility of our ideas. We have been guided by a need to identify critical features of a prototype system that would be reasonably *modular* and *extensible*, and that could be added incrementally to an existing system, such as MEDLARS. The key features of a prototype system of the sort we envision would include all of the following:

- An explicit knowledge base/thesaurus over the important concepts in the database;
- A natural-language processor with the ability to generate semantic representations from user-input expressions, compatable with the knowledge base;
- A means of controlling the context of indexing or retrieval, for example, by the use of 'target' or 'topic' case-frames; and
- A user model, capturing any predictable, pragmatic features of user/system interaction, especially, for example, any *standardized* or *regular procedures.*

We concentrated our attention first on isolating a descriptor vocabulary sufficient for a thesaurus covering the experimental domain of *Management of Rheumatoid Arthritis.* Much of our early effort was devoted to collecting and linguistically analyzing terminology in rheumatology. This included gathering lists of terms from medical dictionaries, textbooks, and standardized controlled vocabularies such as MeSH, ICD-9, and SNOMED, as well as from the titles and abstracts of articles in our experimental database. One of our goals in consulting dictionaries was to determine whether there was enough control on definitions to permit automatic generation of knowledge bases. We chose also to study the indexes of textbooks in rheumatology because the expression of key concepts in such indexes is usually given in the form of a noun phrase in which relevant, related concepts are shown in typical linguistic relations. (See our report in Appendix A.2.1.) Finally, and most importantly, we analyzed the terminology actually occuring in titles and abstracts, as we wanted to identify typical 'rhetorical' patterns, and to distinguish non-domain-specific from domain-specific terminology. One of our objectives in this analysis was to construct candidate frames for the most frequent abstract and title types. (See our reports in Appendix A.2.3 and Appendix A.2.3.) All of this work in linguistic analysis informed our efforts in parser design and thesaurus construction.

Our second principal task focused on the problem of identifying the goals and practices of users. We collected examples of queries from several sources, including the National Library of Medicine, and attempted to design a *query frame' in which the underlying assumptions of users were made explicit. (See our report in Appendix A.3.) Finally, we conducted a pilot study of indexing at the National Library of Medicine to identify, among other features of indexing practice, the procedures used in indexing typical articles from our domain. (See our report in Appendix A.4.)

An ideal, fully-automated system might require no indexing at all; rather, it would respond to a user's query by 'reading' all the information it contained and returning just what was relevant. Ideal systems are not yet feasible, so it is important to identify the facets of current systems that are amenable to partial automation that would improve overall system performance. We believe that the use of semantically complex, but linguistically natural expressions will prove to be the best descriptors in any non-ideal system; and that such descriptors can be built automatically from natural-language expressions. Our efforts in our first year of work have demonstrated the possibility of using artificial intelligence techniques and natural-language processing to achieve some of the automation required to build the next generation of indexing/retrieval systems.

3. Knowledge Representation

Knowledge representation figures in virtually any automation of intelligent behavior. For the ultimate goal of the simulation of an expert physician's ability to identify medical topics, the detailed structuring of medical information is absolutely central. Even when restricted to a particular branch of medicine, the design of a knowledge-representation system calls for much reflection. The system must have *breadth of coverage,* since any title of a medical article is likely to have words for concepts from several medical domains—such as anatomy, physiology, and disorders and their management. All these must be represented in order to represent the content of the title. But also, the system must provide *depth of discrimination,* so that detailed topics can be formulated. This combination of size and conceptual complexity presents an unusual research challenge.

From the start, we believed that the initial phase of the MedSORT project called for a conservative approach to knowledge representation: in an application of this magnitude, it would not be appropriate to attempt to create a new and untried system of representation. We chose a well-tested and highly suitable approach to knowledge representation that has been used in artificial intelligence research for many years. In various forms, the *frame-based* approach has not only proved successful in a variety of applications, but its use has developed many systematic principles of knowledge management that we could exploit in the medical domain. (The literature on frame-based representations is now fairly extensive. See, for instance, [4], [21], and Chapter 8 of [20].)

The standard description of the approach invokes the idea of a *semantic network,* a structure consisting of *nodes,* representing concepts, and *links* of various kinds between these nodes, representing semantic relations. Such a structure can be organized around a system of *frames,* which implement the idea of a node with connections to other nodes. Each frame has a unique *name,* the node name, which is usually a word denoting a desired concept. Additionally associated with each frame axe various *slots,* which may be filled by attributes, by numbers or Boolean values, or by pointers to other frames. This last feature endows frame-based semantic networks with the capacity to represent a multitude of *relations* among concepts; in particular, it provides the ability to express various semantical connections. Computational procedures defined over semantic networks give them dynamic qualities; in particular, *inheritance* algorithms enable information to be associated with concepts by searching for the information that attaches to what we want to regard as more general concepts.

Further reasons for our choice of the frame-based approach to knowledge representation concern the following desirable properties exhibited by the method:

- *Modularity:* Individual frames and collections of frames can be built, modified, and tested in an incremental manner as the frame network expands to increase its coverage. Different topic areas can be built separately and then combined; this permits one research team to work on anatomy, for instance, while another concentrates on drugs. This feature of frame-based system is essential to a project such as ours, since the employment of modularity is the only way to organize large systems.

- *Inheritance:* Information need only be represented at the highest level of abstraction where it applies; it is then automatically inherited by all instances of the general concept. This provides the system with efficient and tested procedures for storing and retrieving information and for validating the correctness of the information.

- *Generality:* All relevant domain information, lexical information, inferential capabilities, consistency-checking methods, and natural-language processing connections can be built and represented in the same uniform frame-based format.

- *Extensibility:* Frame systems are open to theoretical and implementational extensions and refinements. In fact, we have engaged in precisely these activities in our implementation of FrameKit (see description below) and in the development of our thesaurus.

Several distinct types of knowledge must be integrated in the nets being used for the MedSORT system. Full-scale information processing ultimately will require topic representation, indexing, retrieval, language processing, and various inferential processes. It is therefore necessary to have a variety of links represented by what fills the slots in the frames. These presently include the following:

- Taxonomic **relations** *(is-a, part-of, ...)*
- Logical relations *(entails, follows, enables, ...)*
- Linguistic relations *{agent, instrument, location, ...)*
- Domain-specific relations (in the RhA domain, for instance, *finding, treatment, ...*)
- **Pragmatic relations** *(elaborates-on, contrasts-with, ...)*
- **Local attributes** (such as *gender, number, ...*)

A frame integrates the knowledge represented by such heterogeneous relations into **a** single package. Taxonomic information is used primarily to guide the internal flow of information—to determine the source of inherited information, whereas the other categories are the repositories of the various kinds of knowledge useful to specific tasks.

Early in the project Jaime Caxbonell designed and implemented a system he has called FrameKit, which is capable of managing large numbers of frames with relative efficiency (see Appendix D.I.2 for the FrameKit Manual and Appendix D.I.I for the actual LISP code). Our central design principles for FrameKit axe that it should be *streamlined* (i.e., faster than other AI langauges), *flexible* (easy to use and understand), and *extensible* (in case we had not implemented all the requisite functionality on the first pass). FrameKit was intended to provide the central structuring of knowledge available to all processes in present and future MedSORT systems (systems such as the parser, the generator, inferencing systems, the window-display manager, knowledge acquisition routines, *etc.)* It is also an important initial step towards our goal of a multi-functional *indexer}s workbench.*

As we explained, FrameKit allows the user to build a semantic network of frames. Each frame is composed of a body of knowledge stored in slots; relations between frames are established by allowing these slots to be filled by (names of other) frames. A frame consists of a central concept (the *head* of the frame), and a set of semantically well-defined *slots* denoting relations and attributes. The knowledge that is stored in the slots of a frame is often not placed there directly, but is obtained by inferential processes (often *inheritance);* typically, information is inherited by more specific frames from ones that are more general. Moreover, *facets* may be attached to FrameKit slots, which allow the user to control the storage and retrieval of knowledge. So far, seven facets have been defined in FrameKit:

1. Value: This stores the canonical representation of the slot *(e.g.,* "salicylic acid" for the chemical-composition slot of the aspirin frame).

2. View: This places a user-imposed perspective on the slot, which may limit the way in which it is accessed by processes *{e.g.,* allowing a clinician to view clinically-relevant information, but masking irrelevant pathophysiological or chemical aspects of a particular disease).

3. Restrictions: These provide user-defined constraints on slot values *{e.g.,* for use in automating consistency checking as new information is added to an existing knowledge base).

4. If-accessed: This enables user-defined LISP code to fire when a slot's value facet is accessed *[e.g.,* to keep usage statistics).

5. If-added: This enables user-defined LISP code to fire when information is added to a slot's value facet *(e.g.,* to propagate that information to other frames in the network where it may prove useful).

6. If-erased: This enables user-defined LISP code to fire when a slot's value facet has information deleted from it *(e.g.,* to implement a network-based truth-maintenance system such as Doyle's TMS. See [11]).

7. If-needed: This enables user-defined LISP code to fire when no information about a slot's value can be found *(e.g.,* to try to compute the value from other information in the network of frames).

The FrameKit system provides for efficient and flexible creation, deletion and modification of frames, slots and facets. It incorporates functionality providing much user control over the interaction of frames, and the control of inferential processes by means of views. Written in FRANZ LISP, it gives the user full access to a rich interactive LISP environment for program development. The way we have applied this software tool for thesaurus construction is explained in Part II Section 5 below.

4. Natural-Language Processing

Automated natural-language processing can perform two primary and crucial functions in the indexing and retrieval of bibliographic information. The first is in *parsing texts* (first titles, but eventually abstracts and full articles) to extract semantic information required for automated or semi-automated indexing. The second is in *processing direct user queries* to the bibliographic indexing/search system, queries that are stated in natural language. In both of these tasks, complex nominal phrases are of paramount importance: descriptions of diseases, treatments, diagnostic procedures and the like. Such complex noun phrases are especially prominent in the *titles of medical articles* themselves; most often, such a title will be a noun phrase, rather than a complete sentence. Therefore, we have focused our first-year parsing efforts on the automated extraction of a *canonical and unambiguous* semantic representation of complex medical noun phrases. We expect that our later work will turn to the problems raised by query forms, rhetorical devices in titles, connectives in abstracts, and the like.

Our initial parsing effort was focused on a sample corpus of 493 titles on the management of rheumatoid arthritis, provided by the National Library of Medicine. This corpus has enabled us to identify a wide spectrum of parsing problems arising in the parsing and representation of representative medical titles; at the same time, the goal of constructing a robust and extensible parser capable of processing a reasonable fraction of these titles has provided us with a practical constraint by which we can measure the progress of our natural-language processing system. We have found that this list of titles contains a variety of parsing problems; hence, it has served not only as a source of exercises appropriate for the initial development of our parser, but it also poses a series of problems of increasing sophistication.

The goal of any natural-language interface is to translate ordinary-language inputs into formal structures that can be processed computationally. Therefore, any such system must have well-defined *target semantic representations* (e.g., statements in a query or command language) and a representation of the basic components of the phrases to be processed *(e.g.,* a list of words, together with instructions showing how these words can contribute to semantic representations).

As described in Appendix B, a large portion of our one-year project has been devoted to the construction of appropriate semantic representations for the biomedical domain. Our parser's working lexicon is actually quite limited in coverage, being confined to the vocabulary of the fifty sample titles we have parsed. It contains for each of its entries information about syntactic category, necessary synonyms, morphology, and a pointer to the meaning representation in the semantic knowledge base. Alternative pointers aie provided for ambiguous words. See Appendix C.5 for details.

It is the process of *parsing*—of assigning a semantic representation not just to single words but to phrases and sentences—that makes natural-language processing a complex and challenging area of artificial intelligence. Mindful of the need to develop a successful parser based on a well-tested model, and yet seeking to apply new research techniques to our problem domain, we have pursued a two-pronged approach to parsing.

The first approach uses *top-down case-frame parsing.* One realization was based directly on the DYPAR-IV system, a top-down case-frame parser that has been applied successfully in other domains, primarily as an interactive natural-langauge parser for query and command interfaces. (For references to the XCALIBUR project, from which many of our natural language-processing tools derive, see [7], [6]). Briefly, this parsing algorithm works as follows:

1. We associate with each frame a DYPAR-I pattern. This is a 'flat' pattern that has no real hierarchical structure, which is used to identify an instance of the frame.
2. Given an English input, the algorithm matches it against the DYPAR-I patterns. When a pattern is matched, the frame associated with it is made active, and an attempt is made to match the other slots in that frame. This process is constrained by positional and structural restrictions placed on the patterns that can fill slots.
3. This process is repeated recursively on the frame contained in each slot until it arrives at a simple DYPAR-I pattern that fills the slot. Because DYPAR-IV is a non-deterministic parser (*i.e.*, it pursues every possible parse until it succeeds or fails) it must consume all input, and therefore it is not capable of returning partial parses.

The output from the DYPAR-IV system is a fully-instantiated case frame representing the meaning of the sentence. This case frame may either be interpreted as a command to the bibliographic system or used to serve as an index to an associated article.

For a more detailed explanation of the parsing algorithm and descriptions of the uses to which parsers of this sort have been put, see Appendix C.2 and [8] and [9].

As a second approach, we have designed and implemented a prototype of a more flexible lexically-driven, *bottom-up case-frame parser.* It is called **RulePar** and is implemented in a FRANZ LISP-based rule language and agenda-structure called **RuleKit**. (The documented program source code and all available documentation is included under Appendix C.4 and Appendix D.3.) The central advantage to bottom-up parsing in processing titles—an advantage that will be especially useful in later applications to abstracts and full texts—is it that it is *not* an 'all-or-none' parser. If *part* of the text can be understood (or if all the parts can be understood but cannot be connected into a coherent whole), one can still perform a large measure of automated indexing or retrieval tasks, preserving much more information than the key-word frequency approach. Thus, though our desired objective is full-text understanding, the graceful fallback position is maximal partial understanding of the text. (This is illustrated in Appendix C.6, where 50 titles are processed completely into meaning representations for indexing or matching.)

RulePar is a lexically-driven parser with two basic operations: *composition* and *instantiation.* As words are read in, the first and simplest composition operation is attempted: assembling fixed phrases, if any are to be found, from the individual lexical items. Next, starting from the 'true[1] lexicon (words or fixed phrases), syntactic rules of composition are used to propose combinations of words into larger units (such as noun phrases and prepositional phrases). We may say that syntax *proposes a combination;* semantics, on the other hand, must *certify* the composition as legal, and it must *refine* the composition by instantiating appropriate case frames.

Let us consider an example. Take general-purpose frames such as these:

```
[*treat
   agent: •physician
   patient: *person
   treatment: *phys-therapy | *drug
   disorder: •disease I *syndrome]

[•person
   age: *number | *range
   gender: *sex
   occupation: *job I *activity
   name: •proper-name]

[•disease
   stage: *time-course
   type: ...]

[•drug ...]
```

Then, when the parser encounters a phrase (title) such as: *Massive Analgesic Treatment for* ***Advanced Rheumatoid Arthritis in Elderly Male Patients*****, the above frames are instantiated** and composed into the frame below:

```
[•treat
  patient: [*person
               age: *upper-range
               gender: *male
               number: *plural]
  treatment: [*drug
               type: ^analgesic
               dosage: *upper-range]
  disorder:  [^disease
               type: *RhA
               stage: *advanced]]
```

In order to clarify the presentation and convey the central ideas, the semantic representation and the complexity of the parsing process have been simplified in the discussion above. For a full description of both, and completely traced examples, the reader is referred to Appendix C.6.

5. MedSORT Thesaurus Work

Though much of our research effort during the first year of MedSORT Project has been devoted to planning and building a system of frames that represent a portion of biomedical knowledge, we feel that it is more accurate to describe the product of this research as a *thesaurus* than as a medical knowledge base. The reason for this choice of terminology lies in the special nature of our research task: the design of a system for indexing and retrieving medical documents. As we point out in Part HI of this report, the expert human indexer may bring to bear domain-specific knowledge of all kinds in actually classifying a document.

For example, take the following title, from the sample list in Appendix C.I: *Femoral Neck Angles in Osteoarthritis of the Hip.* A knowledgeable indexer might infer that a descriptor such as *Radiographic Findings* applies to this title, from the unlikelihood of the measurements being obtained in any other way. However, the kind of knowledge that informs the system of classification itself should be at once less detailed, and at the same time more stable and less subject to change. The system of classification may well incorporate the information that inflammation is a symptom of arthritis, since this is the sort of information that intuitively belongs to the definition of arthritis. It need not incorporate the information that arthritis is more common in females than in males.

Not all the information recorded in the MedSORT knowledge base belongs to the thesaurus. The most important exception is the linguistic information required by the parser, but many other exceptions will arise as the knowledge base is expanded to add information of the sort needed for medical expert systems of various sorts. Thus, the present thesaurus is best viewed as a subsystem of the knowledge base, corresponding to a certain 'core' view of the entire system.

The modularity of FrameKit, alluded to in Part II Section 3, has enabled us to attack the problem of building the thesaurus in a way that corresponds roughly to the breakdown of medical knowledge into subtopics. These subtopics then correspond to major divisions of the subsumption (or *is-a)* hierarchy of the semantic network. We have constructed representations of the following areas of medical knowledge.

- Human Anatomy
- **The Rheumatic Diseases**
- **Immunology**
- **Medical Procedures,** *including:*
    - Medical Treatments
    - Laboratory Methods
- **Substances,** *especially:*
    - Drugs
- **Medical Equipment**

These divisions will be discussed in more detail in several paragraphs **below.**

Additionally, we can also report that we recently acquired the INTERNIST-I **knowledge** base and have begun integrating it into our thesaurus. The INTERNIST-I expert system has been developed over a period of more than ten years at the University of Pittsburgh School of Medicine and Decision Systems Laboratory by a team of scientists and physicians, including Dr. Harry E. Pople, Jr., Dr. Jack D. Myers, M.D., and Dr. Randolph Miller, M.D., and represents an enormous resource of medical knowledge focused specifically on *diagnosis* (see [14]).

We do not list categories from INTERNIST-I as a branch of our thesaurus, **above,** as it currently exists as a separate, non-integrated network of frames. Nevertheless, it greatly extends our ability to represent general medical information. A further comment on the database can be found at the end of this section.

Anatomy. What we have constructed during the first year of the MedSORT Project provides complete coverage of the body's bones, joints, and topographical regions. It contains more than 250 frames, which were hand-crafted using references such as *Gray's Anatomy* ([12]) and *Mosby's Atlas of Functional Human Anatomy* ([2]) as authorities. *Dorland's Illustrated Medical Dictionary* ([10]) was also useful for resolving many questions.

In constructing anatomical frames we have used the the following links and attributes: ťs-a, *part-of, contains, joins, connected-to*> and *symmetry.* All but the last of these are self-explanatory; *symmetry* was needed to help represent the fact that many anatomical parts are paired, coming in left and right versions. Obviously, the semantic network must be organized so that information about, say, the right hand will be inherited from the generic hand.

The anatomy network is divided into sections corresponding to the major systems of the human body. Inside each system our ẇ-a hierarchies follow the classifications given in *Gray's Anatomy.* For instance, the top of the *is-a* hierarchy for joints looks like this:

- Synarthrosis, *including:*
  - —Sutura
  - —Schindylesis
- Amphiarthrosis
  - —Symphysis
  - —Syndesmosis
- Diarthrosis, *including:*
  - —Ginglymus
  - —Trochoides

And here is a typical anatomical frame:

```
(acromio-clavicular_articulation
        (is-a: arthrodia)
        (located-in: shoulder)
        (joins: clavicle, scapula))
```

See Appendix B.I for a fuller discussion.

Disease. This division contains about 100 rheumatic diseases, organized hierarchically. The structuring of knowledge about diseases is much more of a research problem than that of anatomy; for example, it is not obvious which traits of diseases are central for purposes of classification.

We were very fortunate to have the whole-hearted cooperation of Dr. Thomas A. Medsger, Jr., M.D., in this portion of our project. Dr. Medsger is a member of a group of rheumatologists charged with revising the 1983 ARA Disease Classification (see [1]). He has also written texts on arthritis and many research articles and has given much thought on the question of the organization of his subject domain, which he has shared with us.

This portion of the MedSORT thesaurus was mostly hand-crafted in cooperation with Dr. Medsger. It is based on the ARA disease classification and is meant to represent the point of view of a rheumatological specialist, rather than that of a general practitioner. The rheumatic diseases are first classified as *inflammatory, degenerative, infectious, metabolic,* and *rheumatic states associated with neoplasm.* Further classification gives the *is-a* network a depth of five. By classifying the rheumatic diseases as we have, we are able to associate features of diseases with the most general class. Thus, we can associate all of the manifestations of inflammatory diseases with the class of inflammatory diseases instead of with lupus erythematosus, scleroderma, *etc.*

We have associated each disease with its appropriate treatments. Other links that have been considered (besides *is-a)* are *findings, host, causality, coexistence, predisposition, precedence,* and *component-of.* See Appendix B.2 for a fuller discussion of the disease hierarchy.

Immunology. Also as part of the MedSORT Project, Susanne M. Humphrey of the National Library of Medicine developed a knowledge network of concepts for immunology. The problems that she encountered were somewhat different than those we encountered at Carnegie-Mellon University: immunology is inherently a cross-divisional subject area, and, unlike anatomy, much of what is known about immunology is changing, particularly as it relates to rheumatoid arthritis. The immunology division contains about 50 hand-crafted frames. Its coverage emphasizes immunological cells, agents, and processes. A more thorough discussion of the immunology division of the thesaurus, is included in a document prepared by Susanne Humphrey and is submitted as an appendix to this report. The immunology network includes the following links: *associated-process, associated-system, associated-discipline, associated-cell, differentiation-product,* and *differentiation-source.* See Appendix B.4 for a more complete discussion.

Methods. This division has about 100 frames covering laboratory procedures associated with the rheumatic diseases, and, to a lesser extent, general medicine. It has been drawn from the *Dictionary of Rheumatic Diseases* ([l]) and emphasizes methods related to rheumatic disorders. The structure of this part of the thesaurus corresponds to such broad classes of laboratory methods such as Clinical Pathology (hematology, urine studies, etc.), Imaging (plain radiography, computed tomography, and the like). See Appendix B.3.1 for a more detailed discussion.

Substances. Here we have an *is-a* hierarchy of more than 10,000 frames, drawn directly from section D of MeSH. This was done automatically by a program that read the MeSH file, interpreted the tree numbers, and wrote out FrameKit code.

Treatment. In division we constructed about 250 frames. It has been drawn primarily from *Dorland's Medical Dictionary* ([10]), as well as *Current Therapy* ([16]), and the *Primer of Rheumatic Diseases* ([17]). The thesaurus covers medical treatment in general but gives more attention to aspects of treatment that are important to rheumatology. The major divisions of treatment are as follows:

- Rehabilitational Therapy
  - — Physical therapy
  - — Occupational Therapy
- Psycho-social Therapy, *including:*
  - — Aversion Therapy
  - — Behavioral Therapy
  - — Family Therapy
- Medicinal Therapy, *including:*
  - — Antibiotic Treatment
  - — Anticoagulant Therapy
  - —Chemotherapy
- Surgical Treatment, *including:*
  - —Amputation
  - —Excision
  - —Fusion

In the time frame of the project, we have only been able to start considering slots for treatment frames. Some links that we are using presently are *administrator, method, object, dosage, frequency, duration.* See Appendix B.3.2 for a fuller discussion.

INTERNIST-I. In Appendix B.7 we list the approximately 600 *diagnoses* and 4000 *findings* that comprise the core of the knowledge base, as an indication of the detail and breadth of its coverage. A special problem, here, is that the INTERNIST-I knowledge base is not a frame-based representation system. Every concept, no matter how complex or potentially decomposable, is treated as *atomic*. Thus, a finding such as HEART IRAY CARDIAC SILHOUETTE ABNORMAL LOCALIZED BULDGE does not have links to the more general concepts HEART, XRAY, SILHOUETTE, and BULDGE, though we recognize the importance of these concepts in the composition of various possible natural-language expressions of the finding. In our approach, it is imperative that complex concepts be represented as frames in which there are explicit relations to simpler, more primitive ones. Before the INTERNIST-I knowledge base is fully integrated into our thesaurus, we will have to develop a strategy for decomposing and 'reconstituting' INTERNIST-I's atomic concepts in a manner compatible with our existing representations. See Part III Section 1 and Appendix B.8 for further thoughts on the problems of using the INTERNIST-I knowledge base in its original form.

Part III. Conclusions and Recommendations

As required by the *Statement of Work* for this contract, we must provide in this report a statement of our conclusions and recommendations. Many have already been included in the previous sections, and in this section we summarize these conclusions and provide some additional ones. First we present in some detail a general discussion of our results and of the advice we have accumulated during the project period, then we offer some long-term recommendations, and finally in the last two sections we list conclusions and recommendations in a brief, synoptic form.

1. Review of Project Scope and Results

One of the first conclusions we reached, is that the task as laid out by the original RFP for this contract relates closely to *all* the basic problems in the fields of natural-language processing, indexing, and declarative knowledge representation. These complexities are inherent in the task. Moreover, the highly technical nature of much of the biomedical literature—and the dynamic nature of these fields of knowledge in the long term—combine to make knowledge representation a truly formidable project. Nevertheless, by selecting the research goal of parsing selected titles and by applying well-tested techniques in knowledge representation and natural-language processing, we have succeeded in the first year of the MedSORT Project in establishing the feasibility of our approach. As we argue below, the prototype we have built can be extended to yield useful tools for indexing and retrieval. Furthermore, we would like to stress that it will not be necessary to solve all the background problems we have encountered to produce a working system; though, of course, progress on these problems will improve the performance of the system. The key point is that the medical knowledge-representation module—which is the backbone of the entire system—is relatively stable. It will need to be refined and extended, but it will not need to be discarded.

In this section we discuss our progress over the year under the two main headings of natural-language processing and coverage of domain knowledge.

Natural-Language Processing. The reason why even semi-automated indexing of unrestricted textual data subsumes much of the natural-language processing problem is this: meaning must somehow be extracted from text, and in order to do this linguistic information must be *combined with* domain knowledge. Humans perform this task well—but slowly—and not in a totally consistent manner. Key-word indexing, whether frequency- or cluster-based, cannot begin to provide the fine-grained classification required for reliable indexing. Even a superficial reading of lists of titles of publications shows this, and the interactions between treatments. The development of the knowledge base could not enter the realm of clinical treatments in any detail because *treatment* implies knowledge of *diagnosis* and the mechanics of many *methods.* Among treatment methods, aside from the full range of surgical techniques, *drug treatment* plays a very big rôle. Many (non-MeSH) drug databases are available commercially, but with the available resources of the project it was not possible to try to integrate any of these into the MedSORT database owing to the large amount of natural-language processing that would have to be done on such a resource. We would estimate that—as important as it is—this is a processing task is fully equivalent to the whole MedSORT project effort in conception and scale.

In the second place, the understanding of *rheumatic disorders* requires knowledge of *immunology* and of *diagnostic procedures.* Immunology, of course, is a sophisticated scientific study involving the recognition of highly complex interactions between chemical and biological processes. Diagnosis in turn requires *classification* of rheumatoid disorders. In our earlier discussion of thesaurus work (with comparisons to other classification schemes), it can be seen what an intricate task the construction a sound classification system is. It is clear that experts may very well not agree on particular classification schemata, since rheumatoid diseases are especially long running, usually involving deeply systemic difficulties. The Thesaurus Group could have easily spent most of their time on consideration of immunology or on the task of disease classification.

In the third place, *classification* of disorders requires knowledge of *location,* and thus *anatomy.* Obviously, any medical practitioner requires a detailed knowledge of anatomy, and this need is so basic for any kind of knowledge system, the MedSORT Project could also have very profitably concentrated solely on this area. Though we made very substantial progress, we were simply not able in the short time frame we were given to extend coverage of anatomy to the cardiovascular system, the nervous system, the lymphatic sytem, or the endocrine system.

In consequence, then, our work on thesaurus and knowledge-base construction has focused on the essential *initial steps* of the development, because the imposed subject domain was extremely broad. On the other hand, the appreciation of the difficulties and the progress in delineating a methodology for such work has been quite sound, as many sections of this report argue. And the actual knowledge base we have constructed is non-trivial. Much of our first year's effort has been concentrated on questions of *design* rather than only on questions of *content.* From our experience, therefore, we can make a number of observations about the problems involved in developing knowledge bases adequate for use with natural-language processing systems, including matters of *knowledge organization, validation^ knowledge handcrafting,* and generally the need for engineering control and feedback.

(1) *Knowledge Organization.* The first and, in some sense, most critical decision the builder of a knowledge-base must make is how to organize the concepts. The consequences of this decision will be felt in every other phase of activity. It is clear from our experience that *for purposes of natural-language processing,* knowledge-base organization must reflect the organization of concepts in the expected user, not merely the organization dictated by what might be termed 'taxonomic efficiency'. In our knowledge base, this organization is reflected in the choice of our major sub-classifications—into branches over *anatomy, diseases, medical procedures, substances,* and *medical equipment*—as described in our discussion of the MedSORT thesaurus. These divisions correspond roughly to the 'basic-leveP conceptual divisions that physicians bring to the task of diagnosis and treatment: the default assumptions are that *locations* are anatomic; the *objects of inquiry* are medical disorders; the *activities* involve medical procedures; and the *instruments* of activity will include drugs (substances) and medical equipment. To a certain extent, this kind of division maps onto fundamental divisions among *linguistic case relations,* and thus facilitates natural-language access to the knowledge base.

Several important points concerning *relational links* emerge when the first major subclassifications are made. First, the *domain-specific* links that are required *within* any subclassificatory branch are often not useful in expressing relations among concepts *outside* the branch. Thus, relational links such as *connected-to, part-of,* and *connects* that occur under *anatomy,* have little or no use in expressing relations among concepts under *substances, medical equipment,* or *diseases.* We might note that such links as these get their *semantic sense* from the concepts that they relate, even as the concepts themselves get their sense from a combination of the taxonomic structure of the knowledge base and the relational links that connect them. We can also observe that, in our experience, it is not possible to predict the exact domain-specific links that will be needed or useful in expressing relevant relations among concepts within a conceptual subdivision.

Second, the *domain-independent* or general links that axe required will be either taxonomic *is-a* links or links corresponding to *linguistic case-role relations.* Clearly, *is-a* links are necessary in any classification scheme. Links based on linguistic case roles may not be *pre-theoretically* required for the construction of semantic networks, but they afford an efficient and convenient means of translating between network structure and natural-language expressions.

Third, the principal means of expressing different *points of view* will depend on structuring concepts into parallel, semi-independent hierarchies, typically built from *ẁ-a* relations. These will sometimes cut across major sub-classifications.

(2) *Validation.* A significiant problem for any knowledge engineer is validating the information that becomes encoded in a knowledge base. Without validation a knowledge base cannot be used by a wide community. Validation is especially a problem when the knowledge base is vast and is built and up-dated by numerous individuals who may not share expertise, and it may well prove to be an insuperable obstacle in domains where knowledge is changing rapidly.

In our case, we have found that there is a paucity of validated medical-knowledge source material. Medical dictionaries and textbooks include much information that is obsolete or irrelevant, and many repeat 'facts' that are based on poor but widely-cited research. Even so-called 'standardized' resources, such as the classifications of ICD-9, SNOMED, and *The Dictionary of Rheumatic Diseases* can reflect non-scientific biases, such as the need to restrict all classifications to a depth of 5, or to reach a compromise agreement among members of an advisory committee. Only rarely does one encounter a resource such as the INTERNIST-I knowledge base, in which scrupulous attention has been paid to the establishment of independent authority.

Our knowledge base, unsurprisingly, reflects yaring degrees of validity. By far the most extensively documented portion will consist of the INTERNIST-I knowledge base, as modified and augmented in our representation system, when it is fully integrated. The classification of rheumatic diseases—and to a lesser extent, anatomy—has proceeded under Dr. Thomas Medsger's direction, and thus reflects the judgments of an active clinical and academic physician. The classification of substances derived from MeSH and the classification of medical equipment was based on a textbook article ([18]), so both are subject to the sorts of problems we noted above. We regard it to be a major challenge in knowledge-base design to discover convenient methods for recording *sources* (e.g., texts) and authorities *(e.g.,* authors) of information that comes to be encoded in the knowledge base.

In systems with multiple users and interactive knowledge-base modification, there is also clearly a need for something like *knowledge-base self-documentation*—at least in the form of the automatic generation of *definitions* of concepts. In a complete biomedical knowledge base, such a facility could be the basis of a standardized medical terminology. With more advanced interfaces, it would be desirable to provide automatically generated explanations or justifications for information in the knowledge base.

(3) *Knowledge Handcrafting.* Two problems presented themselves in almost every phase of our work on knowledge-base construction: (i) the need for *handcrafting* knowledge and (ü) the need to strike a balance between *detail* and *generality.* The first of these points underscores the problem of developing new systems based on old information; the second, the difficulty in making knowledge explicit.

We have also found that developing knowledge networks requires considerable creative energy. Multiple interrelated taxonomies must be built in a natural and coherent manner. Different sources of information must be integrated; concepts must be homogenized; often new levels of distinctions must be introduced to provide the bridges necessary to accommodate disparately related facts. Above all, it is a time-consuming task that requires close collaboration with domain experts. A good example of the problem is given in the discussion of our work on integrating the INTERNIST-I knowledge base, below.

Two principles guided us in adjudicating the issue of generality versus detail. First, we included enough concepts to insure that every term we might expect to encounter in the parsing of titles in the target domain could be grounded in the knowledge base. This demands breadth of coverage, but places no special requirement on numbers or kinds of links among concepts. One consequence of this principle is the inclusion of an extensive list of drug names under our *substances* branch, with almost no relations other than taxonomic *is-a* links to one another, and no additional links to other paxts of the knowledge base. Second, we included sufficient richness in inter-conceptual relations to insure that all the types of distinctions we encountered in our sample parses (of 50 titles)—including linguistic and domain-specific relations—were represented in the portions of the knowledge base that the parser most frequently accesses. Consequently, portions of the knowledge base dealing with treatments and methods are relatively richly developed. Clearly, our principles are pragmatic, though not unjustifiable given the absence of precedence in this work and the vastness of the knowledge base.

(4) *The* INTERNIST *Experience.* Having had access to a large, structured knowledge base like INTERNIST-I certainly has facilitated our acquisition of knowledge and has saved us a great deal of time in selecting and classifying information. However, some interesting problems arise when trying to integrate an established knowledge base like INTERNIST-I with one being built for purposes other than diagnosis.

At the time that we acquired the INTERNIST-I knowledge base, we had already built a prototype system of rheumatic diseases for our thesaurus. It was immediately obvious that the classification of diseases in the two knowledge bases was strikingly different. Because the INTERNIST-I disease hierarchy had been designed for the purpose of diagnosis in general medicine, it was built from an anatomical viewpoint; diseases are classified according to the primary system they affect. The diseases that we had classified as *rheumatic* fell within three major categories under INTERNIST: *infectious, joint,* and *systemic* diseases. On the other hand, our hierarchy of rheumatic diseases reflected the perspective of rheumatology clinicians and researchers whose primary interest is in identifying disease mechanisms; thus, we classified rheumatic diseases according to *pathophysiology,* subsuming under one heading all the types that appear in INTERNIST under three. We have, consequently, two well-designed systems that are fundamentally incompatible, neither of which we want to reject.

We have several ideas for integrating the two systems. One would be to 'superimpose' one knowledge base on top of another and allow a user to choose alternative 'views' of the system. Thus, a user would have the option of accessing diseases from the point of view of either a specialist—in particular a rheumatologist—or a general internist. But a great deal of work would have to be done in order to make the two systems compatible under such an *alternative-views* mechanism. We have already done some of this work, such as comparing the two systems in order to identify diseases that are contained in one system but not the other.

And we have taken the next step, which is to classify 'missing' diseases appropriately. This was not always straighforwaxd, since, for example, a disease listed in our rheumatic disease hierarchy (but not in INTERNIST-l) could be infectious from a pathophysiological standpoint, but, while primarily affecting joints, might also affect other anatomical systems. Where would it be placed in the INTERNIST-I scheme: under infectious, joint, or systemic diseases? We were forced at times to make arbitrary decisions. Further ahead lie tasks such as checking the disease links of entries in the INTERNIST-I knowledge base to ensure that relationships between these entries and new entries axe represented, checking disease profiles in INTERNIST-I to ensure that all articular manifestations (e.g., joint inflammation, swelling, degeneration) are encoded, and constructing complete profiles of diseases that we add to the INTERNIST-I knowledge base. Dr. Randolph Miller, one of the designers of the INTERNIST-I system, has suggested that this latter task could easily take as much as two man-years of effort. (For a more complete discussion of the problems involved with integrating INTERNIST-I with our knowledge base, see Appendix B.8.)

2. Possible Future Tasks for Related Projects

Turning now to advice that we can give, we have to note first that the number of possible future tasks that could be set out is nearly endless. We feel, however, that the experience we have gained so far in the MedSORT Project makes it possible to establish some priorities, and to present some realistic recommendations. The list of tasks that we shall give here is designed to illustrate both a range of possible projects that we consider to be relevant to the concerns of the National Library of Medicine—and the biomedical community, generally—and also a coherent, integrated approach to the specific problems of managing biomedical information computationally. In essence, we are proposing a program of research that spans the range from theoretical research to direct applications. Our basic research involves the structuring of knowledge representations to facilitate limited inference, to disambiguate natural-language expressions, and to standardize classification according to significant properties. Our suggested applied research involves building an interface between a knowledge base and a bibliographic database, specifically for making classification and retrieval of information more efficient and responsive to research needs.

But we would like to emphasize that the kind of interface we propose (which might be called an Indexer's Workbench) is also a general-purpose tool for artificial intelligence applications in medicine. In principal, it could serve as the front-end interface (t) between expert systems and tutorial programs, (tt) between non-textual databases and natural-language classification tasks, (tit) between research program managers and the work being generated by the subgroups under their direction—in short, between any structured representational system and natural language-defined tasks. The reason we argue for this generality is that the key features of the knowledge-representation schema we are developing are the ability to manage inheritance of values in tangled hierarchies (where, for example, linguistic properties are represented side-by-side with domain-specific classification features), the standardization of reference via grounding in detailed, unambiguous data structures, and the potential to integrate multi-modal knowledge representations.

We also claim that in order to build a richer indexing system *systematically,* the analysis of titles, abstracts, and texts must be automated as much as possible—always allowing human indexers to supervise, augment or correct the automated analyses. A workbench environment is very much needed to make this possible. The reasons for automating the indexing phase—in part or in whole—include: (t) enhancing the systematic character of the classification, (tt) enabling the re-indexing of large portions of the database (say, according to new criteria or new terminology), (ttt) lessening the economic burden of massive manual indexing, and *(iv)* creating the possibility of the eventual construction of a truly encylopedic medical knowledge base for use by multiple experts and expert systems.

In broader terms, the specific projects areas that we want to recommend can be organized under three long-range research objectives: (1) the construction of standardized biomedical thesauri and deeper knowledge bases (including the continued development of knowledge representation theory and its software implementation); (2) the development of rapid, comprehensive, and research-sensitive biomedical indexing procedures, embodied in the form of a medical *indexer's workbench]* and (3) the pursuit of further basic linguistic research that is needed both for the study of organization and for the construction of effective natural-language tools.

These are all large project topics, and it is clear that substantive progress on tasks of such broad scale will require the collaboration of many groups in the accademic and biomedical communities. We can say that all the points to be made under (1) are basic to *any* artificial intelligence applications, be they classification/retrieval systems or expert systems. But we have seen time and again from our discussions with colleagues working on medical expert systems that their work often depends crucially on solving the representation problem. They need—the biomedical community needs—someone to come forward

to take the lead in articulating a standard that could be used across applications.

We believe that the National Library of Medicine should be the institutional leader in sponsoring and disseminating work on standardized biomedical-knowledge representations. There axe also clearly obvious reasons why the National Library of Medicine should also be a leader in work under (2), but again the results would apply to all indexing applications. Whereas (3) may seem to be too academic, all applied research requires a proper and solid theoretical basis, and we contend that this basic linguistic work must be carried out. Perhaps the National Library of Medicine could begin to work out a collaboration with the NSF or with private foundations that could help in the support of such basic research.

Objective 1: Standardizing Biomedical Knowledge Bases. The accomplishment of this truly gigantic—but essential—task must be broken down into several phases, and many subtasks must be shared among collaborative groups. Before any effective sharing can be envisioned, however, there has to be agreement on the *common way* that knowledge representation is done. We propose that work should proceed in several phases:

1. ***Standardizing Biomedical Knowledge Representation.* Our experience shows that any comprehensive codification should take place in consultation with medical experts in several fields. The issue is the delineation of *forms* of medical knowledge and the agreement on the broad *classifications* of objects, processes, concepts, *etc.* into general categories. Once the types of knowledge to be encoded are understood, the structure of the knowledge base and an appropriate knowledge representation language (such as FrameKit) can be designed, implemented, tested, and refined. The success of the design will require the grounding of ideas in computationally tractable representations as well as medically sound ones.**

2. ***Linking Natural Language to Knowledge Representation.* Certain portions of (English) *grammar* have to be connected with the representations in order to automate input and output *via* natural-language interfaces. The form and the extent of the grammar have to be determined. Certainly it is unrealistic to think that the *whole* of English can be accommodated at once, and procedures for construction of grammatical rules have to be laid down so the work on natural-language features can continue over an extended period. An important goal of this work would be to create a general-purpose and 'theory-neutral[1] *computational system.* Initially, such a system would be most appropriate for interfaces to text-processing tasks, but eventually it could be extended to non-textual domains of application as well. Our RulePar system prototypes the more general language system we envision here.**

3. ***Creation of a Biomedical Thesaurus* The major purpose is to create the next generation of thesaurus beyond MeSH, SNOMED, *etc.* This phase would have**

to incorporate the construction of a *dictionary*, though at any one time the lexicon can be more extensive than the definitional part of the terminology database. The taxonomic organization of such a thesaurus should take into account computationl coherence and tractability. But, it must also be a *medically valid* thesaurus. Such a task of insuring validation is of almost national proportions, so again the leadership of the National Library of Medicine would be vital to its success.

4. *Structuring Biomedical Information both Cross-Modally and Cross-Functionally.* Knowledge is not just verbally recorded. The next phase after fixing the basic representation scheme is to work on structuring databases *cross-modally* (combining, say, visual and linguistic types of information) and *cross-functionally* (encompassing the roles that the 'same' knowledge plays when used in different ways). Radiological and NMR images—together with their interpretations—exemplify the union of verbal and non-verbal knowledge we envision in this eventual integration. The advent of very large capacity optical-disk storage of visual information opens up the possibility of extremely useful interconnections between different kinds of archives. Again, this problem leads to many deep research questions in artificial intelligence. But this research has enormous significance for the management of non-textual collections of all kinds, including medical collections.

Objective 2: Developing a Medical Indexer's Workbench. We must emphasize that we do not expect or propose to eliminate the human indexer. But we can hope to multiply the indexer's capabilities by providing computational tools that will access the classification system and will bring linguistic processing capabilities to bear on the publication to be indexed. A workbench of this sort would be, we feel, the most appropriate vehicle for putting many of our research ideas into practice; it also provides an excellent experimental medium for implementing the new techniques incrementally. The current MedSORT work on automated processing of medical titles furnishes the prototype of an indexer's workbench.

Work on the development of such a utility should proceed in several broad phases:

1. *Developing Tools for Knowledge Acquisition.* Well-structured dictionaries and thesauri are necessary for all other parts of the indexer's workbench we envision. One reason is that users need them as road maps for the large databases. Another reason is that other natural-language programs use knowledge bases as background files in processing. But these road maps may be so large and complex in themselves that special tools and techniques are required for their construction—and especially for their incremental revision and correction. The sheer scale of MeSH shows why this is so. This activity also requires writing guidelines for the construction of dictionaries

and thestiuri. We are claiming here that the knowledge-acquisition problem should *begin* with the construction of domain-specific dictionaries and thesauri and with the building of user-interfaces. These tools should be made usable by domain experts themselves with some technical assistance but without requiring programming on their part. Then the further needs of knowledge acquisition and the further tools that have to be built, can be based on these thesauri.

2. *Developing Robust Software for the Parsing of Titles and Abstracts.* This work is initiated and will become more subtle and more useful as the thesaurus work expands. Its purpose is to implement the connections between language and knowledge representation and to provide the means of manipulating representations. The need is to make discrimination more fine grained, and our solution is to use more involved linguistic constructs in a controlled way. An initial step should be to generate *index frames* that not only reflect the information available from the title of a publication but also incorporate information selectively extracted from an accompanying abstract.

3. *Developing 'Topic-Sensitive' Indexing and Retrieval Procedures.* Part of the problem is 'interface' in the sense of interpreting the user's (or author's) statements with respect to the standardized terminology and knowledge representation. 'Relevance' requires becoming sensitive to topic and context. This relates directly to many issues in current artificial intelligence work, and an implementation objective would be to make experimentation with different strategies of evaluation and search practicable. An immediate and very basic goal should be to identify the specific transformations on a natural-language data structure (partial knowledge representation) that preserve 'topicality' relations in a domain.

4. *Incorporating Indexer/Searcher Practice into the Workbench Environment.* It is not reasonable to write software in a vacuum. After basic parsing programs are available, and after the standardization of terminology and knowledge representation has progressed to the point where there is good coverage of a domain, a system must be made available for use in daily indexing activity—and for use in searching as well. Both aspects of input and output will test the system not only for performance but also for ease, correctness, and quality of use. All problems cannot be anticipated in advance, and only actual use will show how to improve the programming. The comprehensiveness and subtlety of the terminology database also requires this testing. In doing so, it would be possible to collect data on what indexers look for as to topic and essential description. In the other direction, it is necessary to evaluate how searchers configure their searches and how they see relations between concepts. Such feedback from practice is bound to change both the thesaurus

and the knowledge base. For purposes of the National Library of Medicine, the use of any indexcr's workbench will also have to be made consistent with procedures of work on INDEX MEDICUS with the aim eventually of replacing the use of MeSH with the completely computer-held thesaurus and natural-language processing.

Objective 3: Making Progress on Basic Linguistic Research. If we can grant that natural-language processing is vital to the general task, we still have to see how progress can be made in solving the difficulties of machine implementation. We feel that the most significant problems facing a language processing system include:

1. *Ambiguity Resolution:* Many words, phrases, and sentences have multiple meanings in different contexts; thus, the analysis must take complex context effects into account. However, the more standardized the domain lexicon becomes, and the fewer usages of the same term to mean different things, the less serious the ambiguity problem becomes. If one of the objectives of a possible *unified medical language project* is to reduce or eliminate ambiguity in future medical texts and reports, then this aspect of automated language processing becomes much less problematical. There is, nevertheless, a problem that the same terms may have quite different meanings in different specialized domains. Therefore, the ambiguity problem will never go away entirely.

2. *Canonicality:* Natural language allows for a multiplicity of ways of stating the same information. In order not to miss relevant information upon retrieval, the internal information must be reduced to canonical form. Again, standardization of medical language would reduce the complexity of this problem, but agreeing on a canonical form is a *further* step in standardization. A major difficulty, however, is that *people* will not naturally use canonical forms. Therefore, processing has to take place to put their input into the canonical form.

3. *Anaphora, ellipsis, etc.:* Some linguistic phenomena, such as fragmentary phrases or sentences *(ellipsis)* or backward references to previous information *(anaphora),* require complex processing techniques which must be faced and resolved. Note, however, these difficulties would not be particularly ameliorated by a standardized medical lexicon, since they concern grammatical structure rather than vocabulary. We are certainly not claiming that software developments utilized in the MedSORT project have gone a non-trivial distance in solving these problems.

Although we are not expecting to solve all of the problems confronting natural-language understanding systems, we do see the possibility in our research of extending automated processing to much more complete forms of natural-language text than is possible through what we know to be the current approaches. We have designed a parser that is capable of utilizing discourse/pragmatic information in addition to syntactic information to produce semantic representations directly. And the semantic representations (i.c, the *case frames)* are compatible in their structure with the knowledge representations that lie at the heart of the system.

A significant feature of the approach we have taken is its ability to utilize identical data structures for knowledge representation and for language generation, providing the direct link needed in any genuinely interactive interfacing with a user. Another feature is the possibility inherent in the representational schemata to present similar information from different 'points of view', providing the basis for treating notions like *topic* computationally and for managing the problems associated with *vagueness*—as when different users employ the same terms with different degrees of precision or with differing senses.

3. Summary of Conclusions

We repeat here in a brief form the chief *conclusions* that we have reached in our first year of research into automated bibliographic information retrieval in the medical domain. The next section summarizes *recommendations.*

- The system we have designed and its implementation demonstrate the *feasiblity* of combining breadth of coverage in the biomedical domain with depth of knowledge representation and efficient parsing.
- *Frame-based systems* of knowledge representation axe well suited to the purpose of *codifying* biomedical information to be used in *indexing* and in *natural-language processing.*
- *Conventional resources* such as dictionaries and standardized nomenclatures provide helpful information, but they *cannot* be automatically converted into a well-designed semantic network.
- In large multi-user information-processing systems it is imperative to enforce *internal standardization of the structures that encode biomedical concepts.* Semantic-based *canonical* representations overcome the obstacles of ambiguity and variability in the interpretation of natural-language expressions.
- Natural-language processing offers the only useful *automated* technique for converting free text and queries into cannonical representations that axe suffciently detailed in *depth* and *breadth* to span large, real-world domains.
- Interactions between a system of *knowledge representation* and a *natural-language parser* are too complex to allow components of a system to be developed indepen-

dently, especially if the systems are extensive in their coverage.

- A *combination* of approaches to the parsing problem makes the necessary processing quite feasible. A *bottom-up case-frame* parser (such as RulePar) can be used to guarantee that *partial* representations can be generated in the face of uncertainty. A *top-down case-frame* parser (such as DYPAR-IV) can be used to control the *contexts* under which partial information is to be interpreted.
- One of the most natural applications of combining detailed knowledge representation with natural-language processing is in the *semi-automated indexing* of free text.
- Long-range success in projects in automated, intelligent medical indexing is heavily dependent on a research environment in which useful *tools* can be developed and tested. Many of these tools would also be useful in other areas of medical artificial intelligence.
- Cooperation with medical experts is essential both for research and development in biomedical informatics.

The principal investigators trust that these conclusions have been sufficiently justified in this report and its accompanying documentation. We are confident that our work during the past fifteen months has faithfully and fully addressed the research interests of the National Library of Medicine as described the original RFP.

4. Summary of Recommendations

Aside from the obvious recommendation that much more work in this area is needed, we can summarize our general recommendations as follows:

- The results of our first year provide the basis for developing an *indexer 's workbench.* Research, in this direction, guided by a series of incremental goals, should aim at creating interactive tools that will enhance the ability of a human indexer to classify documents. Work on knowledge representation, on refining and broadening coverage of the databases, and on the development of the parser are naturally included in the context of this project.
- The development of the next generation of biomedical information-management systems—including standardization of knowledge representation—requires intensive collaboration between medical experts and computer scientists. The National Library of Medicine should work to promote collaborative research teams wherever possible.
- The whole field of Biomedical Informatics needs direction and leadership. The

National Library of Medicine is the best-positioned Federal agency to assume such leadership; they have led in the past and should continue to do so in the face of evolving technologies.

We believe that the development of computational resources for biomedical information management should be pursued as quickly as possible. Whereas university laboratories—such as our own at Carnegie-Mellon or that of the University of Pittsburgh—can play a crucial role in the basic research, only a Federal agency whose mandate is the codification, structuring, use and maintenance of medical knowledge can take the lead in long-term and large-scale projects of national interest.

For our own part, the MedSORT group are prepared to undertake the following specific tasks:

1. Refine the *structuring principles of biomedical-knowledge* representation based on sound epistemological methods and accurate biomedical knowledge.
2. Develop a means of *managing different categories of knowledge* that share common aspects *(e.g.,* clinical, pathophysiological, pharmaceutical, biochemical, radiological) in the same integrated knowledge base.
3. Modify and *augment* FrameKit to embody these refinements and additions in a computationally effective manner.
4. Provide *knowledge-acquisition tools,* so that domain experts *[e.g.,* physicians and other specialists) can augment the knowledge base directly without requiring an intermediary.
5. Develop further a *extensible and comprehensive thesaurus* to allow incorporation of new, more detailed biomedical knowledge bases.
6. Begin exploration of a *complete subdomain* of biomedical knowledge *[e.g..* rheumatology, oncology, or neuropathology) to determine in what ways the indexing and retrieval provided by comprehensive knowledge structuring is significantly superior to the MEDLARS system.
7. Augment our *lexically driven parser* so that it is able to cope with larger *syntactic variation* and to extract *maximal information* from even *partial understanding* of more complex titles and abstracts.
8. Develop an XCALIBUR-like *user interface* in natural language for users to be able to formulate their retrieval queries.
9. Add a *discourse component* to the natural-langauge facilities so that higher-order relations can be extracted from texts *{e.g.,* abstacts), and used for more accurate and finer-grained indexing.

Our intention is to continue collaborating with more than one group of medical experts. A very long-term objective is to build *high quality, extensible expert systems* on top of our knowledge base, and we see the proposed continuation of the MedSORT work as an essential step in that direction. In particular, would propose to cooperate more closely with the group of Computer Scientists and Physicians involved in the CADUCEUS project at the University of Pittsburgh. The connections we have built up have been extremely encouraging and productive. The momentum and team work must be maintained.

Part IV. Project Activities

During this first year of MedSORT work, we organized or participated in several extra-research activities that sharpened our understanding of the issues in our project. Though we have reported on some of our activities previously, we collect in appropriate appendices all our earlier reports as well as material that is new with this final report. We also append a list of these activities here.

(1) In December 1984, we organized two weekend workshops for our collaborators at the Lister Hill National Center. The first was an orientation/training session on our project's parsers, DYPAR-I and DYPAR-IV. The second was a special session on issues in medical thesaurus construction. See Appendix E.I and Appendix E.2, respectively, for our reports on these activities.

(2) In January 1985, we sponsored a seminar by Dr. Martin Kay of the XEROX Palo Alto Research Center on lexicology issues.

(3) In February 1985, several members of the MedSORT Project team visited the Lister Hill National Center to offer tutorials in frame representation. See Appendix E.3.

(4) In March 1985, we hosted a three-day weekend workshop on thesaurus construction, with visiting computer scientists from Bell Communications Research, information scientists from the University of Pittsburgh and Carnegie-Mellon University, and medical experts from the Decision Systems Laboratory at the University of Pittsburgh School of Medicine. Along with a copy of our earlier report on this activity, we include a copy of our background notes on the meeting and transcripts of three presentations made during the meeting by members of the MedSORT Project Team. See Appendix E.4.

(5) In April 1985, we presented a six-month progress report to the National Library of Medicine. Our agenda and copies of the slides from that report are in Appendix E.5.

(6) Finally, in July 1985, we made a presentation on the MedSORT Project to the Board of Scientific Counselors of the National Library of Medicine. Copies of our notes and slides from that presentation, along with copies of the Board's review, are included in Appendix E.6.

(7) In addition to the scheduled activities listed above, the MedSORT Project team has been developing close ties and collaborative activities with members of the University of Pittsburgh School of Medicine. One of our many cooperative projects has led to the inclusion of the uniquely detailed and authorative INTERNIST-1 knowledge base in our thesaurus. We have also initiated a weekly seminar on issues in Artificial Intelligence and Medicine.

Part V. A Guide to the Appendices

1. Table of Contents

A. Indexing Issues

B. Knowledge Representation

C. Natural Language Processing

2. A Precis of the Appendices

A. Indexing Issues

A.I. A Review of Three Indexing Methods

The indexing systems MEDLARS, PRECIS, and SMART are described and then evaluated in terms of the following criteria: exhaustivity of indexing, specificity of indexing, recall in retrieval, precision in retrieval, and consistency of indexing. The potential contribution of work in artificial intelligence systems is discussed.

A.2. Analyses of Linguistic Data

A.2.1. Conventional Sources of Domain Knowledge

This report describes one of the first activities of the Natural-Language Group: an exploration of the conventional sources of medical knowledge. Medical reference books and medical dictionaries were examined to determine the terminology they contained and the various ways this terminology is used. The results of this examination were used as **a** basis for future research into the development of a user interface and an indexing system. The implications of these results for the construction of a retrieval system are also discussed.

A.2.2. An Analysis of Entry MeSH Headings and Subheadings

An exploratory analysis was made of the use to which subheadings are put in the heading-subheading pairs characterizing entries. The short-term goals of this investigation were to (f) study the general syntactic and semantic nature of the subheadings, and *(it)* to compare them to terms found in indexes of medical sources such as the *Primer on the Rheumatic* ***Diseases.***

1. Subheadings' Frequencies and Co-Occurrences
2. Subheadings Definition Structure

A.2.3. An Analysis of Entry Titles

A file of 493 sample bibliographic entries, bearing on drug therapy of rheumatoid arthritis and the diagnosis of selected rheumatic diseases, were made available for analysis (see Appendix C.I). This report investigates: (i) titles' lexicon and its relationship to the specialized lexicon provided by the AFP and RDDM indexes that we analyzed (see Appendix A.2.1); *(ii)* how the lexicon might be organized so as to aid grammar-writers in detecting regularities; ***(tit)*** syntactic, semantic, and pragmatic aspects of the titles that would either allow or pose serious problems for parsing.

1. 493 Sample Titles
2. Words in Titles (Alphabetical Order)
3. Words in Titles (Frequency Order)
4. Words in Titles but not Abstracts
5. Words in Textbook Indexes (Alphabetical Order)

6. Words in Textbook Indexes (Frequency Order)
7. Words in Titles but not Textbook Indexes
8. Titles Containing Potentially Rhetorical Terms
9. Notes on Rhetorical Words in Titles
10. Stop Words
11. 2, 3, and 4-Word Phrases in Titles and Abstracts
12. Preliminary DYPAR Title Grammar
13. Segmented Titles

A.2.4. An Analysis of Entry Abstracts

On the basis of our analysis of the structure of abstracts, it seems likely that some predictably frequent structures can be identified. These regular abstract-forms will aid both in the disambiguation of terms that appear in titles, and in the identification of key concepts in articles.

1. Words in Sample Titles and Abstracts (Alphabetical Order)
2. Words in Sample Titles and Abstracts (Frequency Order)
3. Words in Abstracts but not Titles
4. Words' Propensities to Appear only in Abstraces
5. Abstracts' Frames

A.3. An Analysis of User Queries

The Natural-Language Group analysed user queries to inform the design of both a user interface and an indexing system. The results of a preliminary analysis show that queries exhibit features which would be difficult or impossible for the natural language parser of a user interface to handle. Queries were also shown to contain other features that must be accomodated in the design of an indexing system. Several conclusions based on this preliminary analysis of queries are presented here. The need for future, more rigorous analysis of queries is also discussed.

1. Falk Library Queries

A.4. A Protocol Study of Indexers at the NLM

This report contains the results of a pilot protocol study of indexers at the National Library of Medicine. The three purposes of the study were: to determine the effectiveness of protocol analysis in studying indexing methods, to generate hypotheses on which future studies might be based, and to evaluate the testing methodology used in the study. The study revealed that the protocol method was indeed effective for analyzing indexing methods. Based on the analysis of the protocol data gathered, several hypotheses concerning indexing methods were generated. Several recommendations for the design of future protocol studies are also presented.

B. Knowledge Representation

B.I. Human Anatomy

Anatomy is one of the most important divisions of medical knowledge, and the one that we chose to represent first. The system contains approximately 250 frames and is fairly complete for the articular and skeletal systems and topography.

1. Representing Human Anatomy
2. Database on Human Anatomy

B.2. Rheumatic Diseases

Since our prototype domain is rheumatology, a good classification of the rheumatic diseases is crucial for our system. This network was culled from the *American Rheumatism Association Disease Classification.* It represents all diseases known to rheumatologists and meets an expert's standards for usability.

1. Representing the Rheumatic Diseases
2. Database on Rheumatic Diseases

B.3. Medical Procedures

B.3.1. Laboratory Methods and Findings

like other fields of medicine, rheumatology contains a well-defined set of laboratory methods and findings. Our network is based on the *Table of Contents* to Volume II of the *Dictionary of the Rheumatic Diseases.*

1. Representing Laboratory Methods and Findings
2. Database on Methods and Findings

B.3.2. Medical Treatments

We represent medical treatment in general, but handle in greater depth aspects of treatment that are important to rheumatology, and in particular physical and occupational therapy. The network is primarily handcrafted, containing about 250 frames. This represents our first attempt to build a *large* on-line database from the ground up. We were forced to do this because information was not available from MeSH or from INTERNIST-I.

1. Representing Treatments
2. Database on Treatments

B.4. Immunology

This representation of the human immune system is based on the following links tš-a, *assoc-cell,assoc-process,assoc-discipline,part-of,differentiation-source,*and*differentiation-product.*The frames in this part of the knowledge base were handcrafted by Susanne Humphrey of the National Library of Medicine.

B.5. Substances

Substances (chemicals and drugs) are an important part of the medical domain and thus should occupy a major section of a medical thesaurus. We applied our FrameMaker program to section D of MeSH, resulting in a network of over 10,000 frames.

B.6. Medical Equipment

Our network of therapeutic devices is restricted to rheumatology, but its structure favors expansion. We handcrafted the network using an article on arthritis therapy.

1. Representing Medical Equipment
2. Database on Equipment

B.7. INTERNIST Knowledge Base Materials

We have incorporated the INTERNIST-I knowledge base of manifestations and diagnoses into our system, with the permission of Drs. Randolph Miller, Jack D. Myers, and Fred Masarie of the University of Pittsburgh School of Medicine and Decision Systems Laboratory. The list of manifestations contains approximately 4000 entries; the diagnosis knowledge base contains approximately 600 entries.

1. Alphabetical Listing of Diseases
2. Hierarchical Listing of Diseases
3. Alphabetical Listing of Manifestations

B.8. INTERNIST-I and the RhD Classification

This appendix contains materials that are being used to integrate the Rheumatic Disease Hierarchy that we designed into the INTERNIST-I Disease Hierarchy. In particular, they present a comparison between the two knowledge bases, focusing on the similar and distinct components of each. In addition, they present a preliminary attempt to classify diseases contained in the Rheumatic Disease Hierarchy in the context of the INTERNIST-I hierarchy, for items missing from the latter.

1. Overview
2. ARA Diseases Not in INTERNIST-I (1)
3. ARA Diseases Not in INTERNIST-I (2)
4. Rheumatic Diseases Not in ARA
5. Diseases Associated With Rheumatic States
6. Suggestions for Revising INTERNIST-I
7. Placement of ARA Diseases in INTERNIST-I

B.9. References

This section contains the list of references cited in Appendix B.

C. Natural Language Processing

C.I. A Listing of 493 Titles

This section contains a listing of 493 titles of articles about rheumatology. The natural language processing effort within the MedSORT project has been driven in large part by reflection on these titles.

C.2. DYPAR-I

DYPAR-I is a rule-based natural-language interpreter, adaptable to many limited domain applications such as database query, command interpretation for software systems, and simple expert system command and query tasks. DYPAR-I is designed to serve as a high-level programming tool, making it possible for anyone to write a grammar for a specific application. DYPAR-I is implemented in FRANZ LISP and runs under VAX UNIX or VAX EUNICE. This interpreter is a proper subset of other, more powerful parsers of the DYPAR family.

C.3. DYPAR-V Source Code

DYPAR-V is a more powerful member of the DYPAR family than DYPAR-I and is still under development. A partial listing of its source code is provided here.

C.4. RulePar

RulePar is a lexically-driven CF parser that handles the following natural language processing phenomena: compoimd nouns having any number of nouns, adjective-noun compounds having any number of adjectives before any number of nouns, determiners and quantifiers at the beginnings of noun phrases, prepositional phrases and PP-attachment, and fixed phrases. The parser operates primarily in a bottom-up form; its output consists of instantiated FrameKit frames that may point to other frames corresponding to subordinated constituents.

C.5. The Lex Files

This part of Appendix C contains a lexical database for parser testing, including sample inputs, articles, conjunctions, quantifiers, and nouns (disease names, drug names, body-part names, and higher-level concept names).

C.6. Sample Parser Runs

This file is a set of working inputs for the RulePar parser together (with parses), together with traced parses for four titles. The sample inputs listed here are a very small subset of the phrases that will currently work with the parser.

1. 50 Titles
2. Parses
3. Traced Parses
4. Traced Partial Parses

C.7. A Fragment of the MedSORT Thesaurus

This section contains a listing of the fragment of the MedSORT thesaurus that is used by the parser.

C.8. KAFKA

Many LISP programs transform one S-expression into another. The KAFKA language and the KAFKA interpreter provide a clear way to specify transformations for a subset of S-expressions called *case frames* using *Match and Instantiate* rules. This transformational capability is a powerful way to encode linguistic knowledge in the form of production rules.

D. Utility Programs

D.I. FrameKit

FrameKit is a knowledge representation language that enables the user to define relationships between information-bearing structures called *frames.* FrameKit also enables the user to store and retrieve information using three kinds of inheritance that exploit the **is-a** link.

D.2. Frame Building Utilities

This section contains source code for three frame building utilities.

D.2.1. FrameMaker

FrameMaker is a C-shell program that converts indented text to frame form.

D.2.2. NLM-FrameMaker

NLM-FrameMaker converts MeSH data to frame form.

D.2.3. DeFramer

DeFramer is a LISP function that converts frames to indented text.

D.3. RuleKit

This section of Appendix D contains source code for RuleKit, a rule-generating utility that is useful in the design of expert systems.

E. Project Activities

E.I. Workshop on DYPAR

This section contains notes from a workshop on the DYPAR family of parsers.

1. Report on DYPAR tutorial
2. Agenda of DYPAR tutorial
3. Note on DYPAR tutorial

E.2. MedSORT Thesaurus Workshop

This section contains notes from a workshop on medical thesaurus construction.

1. Report on Thesaurus Construction Workshop

2. Agenda of Thesaurus Construction Workshop

3. Note on Thesaurus Construction Workshop

E.3. Frame Representation Tutorial

This section contains the slides from a frame-representation tutorial conducted by John Aronis and Sandra Katz at the National Library of Medicine. Also included is a preliminary report on representing human anatomy for medical information retrieval.

1. Report on Frame Representation Tutorial

2. Slides from Frame Representation Tutorial

E.4. Thesaurus Weekend Workshop

The Thesaurus Workshop was held on March 1, 2, and 3, 1985. Notes from this workshop, together with addresses by Dana S. Scott, David A. Evans, and Jaime G. Carbonell, are included here.

E.5. The Six-Month Report to the NLM

This section contains the slides that accompanied the report that was submitted by the members of the MedSORT Project to the National Library of Medicine on April 29, 1985.

1. Agenda of Six-Month Report

2. Contract-Program Review (Slides)

E.6. The BoSC Presentation

This section contains excerpts of the presentation by the members of the MedSORT Project to the National Library of Medicine Board of Scientific Counselors on July 17, 1985.

1. Comments of the BoSC

2. Slides from BoSC Presentation

Part VI. Bibliographies

References

The following sources are cited in the body of this Report. For a more complete sampling of the relevant literature, see the General Bibliography that follows this one.

[I] American Rheumatism Association, *Dictionary of the Rheumatic Diseases: Volume II: Diagnostic Testing.* Contact Associates International, Ltd.: New York.

[2] Beck, E.W., *Mosby's Atlas of Functional Human Anatomy.* Mosby, 1982.

[3] Blair, D. C. and Mar on, M. E., "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System." *Communications of the ACM 28,* (1985), pp. 289-299.

[4] Brachman, R.J. and Levesque, H.J., *Readings in Knowledge Representation.* Morgan Kaufmann, Los Altos, 1985.

[5] Caxbonell, J. G. "Towards a Self-Extending Parser", *Proceedings of the 17th Meeting of the Association for Computational Linguistics* (1979), pp. 3-7.

[6] Caxbonell, J.G., Boggs, W.M., Mauldin, M.L. and Anick, P.G.,"The XCALIBUR Project, A Natural Language Interface to Expert Systems," in *Proceedings of the Eighth International Joint Conference on Artificial Intelligence,* 1983

[7] Caxbonell, J.G., Boggs, W.M., Mauldin. M.L.. and Anick, P.G., "The XCALIBUR Project, A Natural Language Interface to Expert Systems and Data Bases," in S. Andxiole (ed.), *Applications in Artificial Intelligence,* (Petrocelli Books Inc.), 1985.

[8] Caxbonell, J. G. and Hayes, P. J. [1983] "Recovery Strategies for Parsing Extragrammatical Language," *American Journal of Computational Linguistics,* vol. 9, no. 3-4, pp. 123-146.

[9] Caxbonell, J. G. and Hayes, P. J. "Natural Language Processing: Techniques and Applications," in *The Encyclopedia of Artificial Intelligence,* Shapiro, S. (editor). Wiley & sons, New York, NY, 1986.

[10] *Dorland's Illustrated Medical Dictionary,* 26^{tA} ed. W. B. Saunders Company: Philadelphia, Pennsylvania, 1985.

[II] Doyle, J. "A Truth Maintenance System," *Artificial Intelligence,* vol. 12, no. 3, 1979.

[12] Gray, H. *Anatomy of the Human Body, 13th* American ed. Edited by Carmine D. Clemente. Lea & Febiger: Philadelphia, Pennsylvania, 1985.

[13] Hayes, P. J. and Mouradian, G. V. "Flexible Parsing," *Proceedings of the 18th Meeting of the Association for Computational Linguistics,* (1980), pp. 97-103.

[14] Miller, R.A., Pople, H.E., and Meyers, J.D., "INTERNIST-I, An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine." *New England Journal of Medicine,* vol. 307 (1982), pp. 468-476.

[15] Perrault, C. R., Allen, J. F. and Cohen, P. R. "Speech Acts as a Basis for Understanding Dialog Coherence," *Proceedings of the Second Conference on Theoretical Issues in Natural Language Processing,* 1978.

[16] Rakel, Robert E., ed. *ConnJs Current Therapy: Latest approved Methods of Treatment for the Practicing Physician.* W. B. Saunders Company: Philadelphia, Pennsylvania, 1985.

[17] Rodman, Gerald P., *et al. Primer on the Rheumatic Diseases, 8th* ed. Arthritis Foundation: Atlanta, 1983.

[18] Swezey, Robert L. "Rehabilitation Aspects in Arthritis," in D. J.McCarty, ed., *Arthritis and Allied Conditions.* Lea & Febiger: Philadelphia, 1985.

[19] Schank, R. C. *Conceptual Information Processing,* Amsterdam: North-Holland, 1975.

[20] Winston, P.H. *Artificial Intelligence, 2nd* ed. Addison-Wesley: Reading, Massachusetts, 1984.

[21] Woods, W. A. "What's in a link: Foundations for Semantic Networks." In D. J. .Bobrow and A. M. Collins (eds.) *Representation and Understanding: Studies in Cognitive Science,* Academic Press: New York, 1975, pp. 35-82.

General Bibliography

The following bibliography contains **a** sampling of the literature that formed the background to our work. It is not meant to be exhaustive of any of the fields represented.

Anatomy and Physiology

BECK, E.W., **and GROER, M.**

1982 Mosby's atlas of functional human anatomy. The C.V. Mosby Company, 1982,

FRIEL, J.P., ***et al., editors***

1985 Borland's illustrated medical dictionary ***(26th*** **edition). W.B.Saunders Company, 1985.**

GRAY, H.

1985 Anatomy of the human body (13t/l American edition). Edited by Clemente, CJ).. Lea and Febiger, 1985.

Information Retrieval and Thesaurus Construction

m **AUSTIN, D., and DYKSTRA, M.**

1984 PRECIS ***{2nd*** **edition). The British National Library, 1984.**

BRITISH NATIONAL LIBRARY.

1984 *Guidelines for the establishment and development of monolingual thesauri.* **Technical report. Draft revision of BS5723, 1984, 66 pp.**

RICHMOND, P.A.

1981 Introduction to PRECIS for North American usage. Libraries **Unlimited, Inc,1981.**

S ALTON, M., and Me GILL, M.J.

1983 Introduction to Modern Information Retrieval. McGraw-Hill Book Company, 1983.

SOERGEL, D.

1974 Indexing Languages and Thesauri: Construction and Maintenance. Melville Publishing Company, 1974.

1985 **Organizing Information: Principles of Data Base and Retrieval Systems. Academic Press, 1985.**

Knowledge Representation

AlKINS, J.S.

1984 *A representation scheme using both frames and rules.* **In: Rule-Based Expert Systems, edited by B.G. Buchaman, and E.H. Shortliffe. Addison-Wesley, 1984, pp. 424-440.**

BOBROW, D.G., and WINOGRAD, T.

1977 *An overview of KRL, a knowledge representation language.* **Cognitive Science, vol. 1(1) (1977), pp. 3-46.**

Reprinted in Brachman and Levesque 1985.

BRACHMAN, R.J.

1979 *On the epistemological status of semantic networks.* **In: Associative Networks: Representation and use of knowledge by computers, edited by N.V. Findler. Academic Press, 1979, pp. 3-50.**

Reprinted in Brachman and Levesque 1985.

1983 *What IS-A Is and Isn't: An analysis of taxonomic links in semantic networks.* **Computer, vol. 16(10) (1983), pp. 30-36.**

1985 *"I lied about the trees'^{9} or, defaults and definitions in knowledge representation.* **The AI magazine, Fall (1985), pp. 80-93.**

BRACHMAN, R.J., FIKES, R.E., and LEVESQUE, H.J.

1983 *KRYPTON: Integrating terminology and assertion.* **In: Proceedings of AAAI-83. 1983, pp. 31-35.**

1983 *Krypton: A functional approach to knowledge representation.* **Computer, vol. 16(10) (1983), pp. 67-73.**

Reprinted in Brachman and Levesque 1985.

BRACHMAN, R.J., GILBERT, V.P., and LEVESQUE, H.J.

1985 *An essential hybrid reasoning system: knowledge and symbol level accounts of Krypton.* **In: Proceedings of the ninth International Joint Conference on Artificial Intelligence. 1985, vol. 1, pp. 532-539.**

BRACHMAN, R.J., and LEVESQUE, H.J.

1982 *Competence in Knowledge Representation.* In: Proceedings of AAAI-82. 1982, pp. 189-192.

1984 *The tractability of subsumption in frame-based description languages.* In: Proceedings of AAAI-84. 1984, pp. 34-37.

BRACHMAN, R.J., and LEVESQUE, H.J., *editors*

1985 Readings in Knowledge Representation. Morgan Kaufmann, 1985.

BRACHMAN, R.J., and SCHMOLZE, J.G.

1985 *An overview of the KL-ONE knowledge representation system.* Cognitive Science, vol. 9(2) (1985), pp. 171-216.

CARBONELL, J.G.

1978 *POLITICS: Automated Ideological Reasoning.* Cognitive Science, vol. 2 (1978), pp. 42-46.

DOYLE, J.

1979 *A Truth Maintenance system.* Artificial Intelligence, vol. 12(3) (1979).

DREYFUS, H.L.

1981 *From micro-worlds to knowledge representation: AI at an impasse.* In: Mind Design, edited by J. Haugeland. The MIT Press, 1981.
Reprinted in Brachman and Levesque 1985.

FAHLMAN, S.E.

1979 NETL: A System for Representing and Using Real-World Knowledge. The MIT Press, 1979.

1979 *Representing and Using Real- World Knowledge.* In: Artificial Intelligence, an MIT Perspective, edited by P.H. Winston and R.H. Brown. The MIT Press, 1979.

FIKES, R., and KEHLER, T.

1985 *The role of frame-based representation in reasoning.* Communications of Association for Computing Machinery, vol. 28(9) (1985), pp. 904-920.

FININ, T., and SILVERMAN, D.

1984 *Interactive classification: A technique for building and maintaining knowledge bases.* In: Proceedings of the IEEE Workshop on Principles of Knowledge-Based Systems. 1984, pp. 107-114.

HAYES, P.J.

1974 *Some problems and non-problems in representation theory.* In: Proceedings of AISB Summer Conference. 1974, pp. 63-79.
Reprinted in Brachman and Levesque 1985.

1979 *The logic of frames.* In: Frame Conceptions and Text Understanding, edited by D. Metzing. Walter de Gniyter and Company, 1979.
Reprinted in Brachman and Levesque 1985.

GARY, G.G.

1979 *Encoding Knowledge in Partitioned Networks.* In: Associative Networks: Representation and Use of Knowledge by Computers, edited by N.V. Findler. Academic Press, 1979.

LEVESQUE, H.J.

1984 *A fundamental tradeoff in knowledge representation and reasoning.* In: Proceedings of CSCSI-84. 1984, pp. 141-152.
Reprinted in Brachman and Levesque 1985.

MASARIE, F.E., MILLER, R.A., and MYERS, J.D.

1985 *Internist-I properties: Representing common sense and good medical practice in a computerized medical knowledge base.* In: Computers and biomedical research. Academic Press, 1985.

MCCARTHY, J.

1977 *Epistemologiccd problems of artificial intelligence.* In: Proceedings of the fifth International Joint Conference on Artificial Intelligence. 1977, pp. 1038-1044.
Reprinted in Brachman and Levesque 1985.

MCDERMOTT, D.

1976 *Artificial intelligence meets natural stupidity.* SIGART Newsletter, no. 57 (1976), pp. 4-9.

1978 *Planning and Acting.* Cognitive Science, vol. 2 (1978), pp. 71-109.

MILLER, R.A., POPLE, H.E., and MEYERS, J.D.

1982 *INTERNIST-I, An Experimental Computer-Based Diagnostic Consultant for General Medicine.* New England Journal of Medicine, vol. 307 (1982), p. 468-476.

REIGEIt, C.

1975 *The Commonsense Algorithm as a Basis for Computer Models of Human Memory, Inference, Belief and Contextual Language Processing.* In: **Proceedings of TINLAP-I. 1975.**

SCHANK, R.C.

1975 *Conceptual Dependency Theory.* **In: Conceptual Information Processing. North-Holland, 1975.**

1980 *Language and Memory.* **Cognitive Science, vol. 4(3) (1980), pp. 243-284.**

SCHMOLZE, J.G., and LIPKIS, T.A.

1983 *Classification in the KL-ONE knowledge representation system.* In: **Proceedings of the eighth International Joint Conference on Artificial Intelligence. 1983, vol. 1, pp. 330-332.**

SOWA, J.F.

1984 Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley Publishing Company, 1984.

WINSTON, P.H.

1984 Artificial Intelligence. Addison-Wesley, 1984.

WOODS, W.A.

1975 *What's in a link: Foundations for semantic networks.* **Representations and Understanding, edited by D.B. Bobrow, and A. Collins, editors. Academic Press, 1975.**

Medical Science

AMERICAN RHEUMATISM ASSOCIATION GLOSSARY COMMITTEE.

1982 Dictionary of the rheumatic diseases. (Vol. I: Signs and Symptoms). Contact Associates International Limited. 1982, ii+95 pp.

1982 Dictionary of the rheumatic diseases. (Vol. II: Diagnostic Testing). Contact Associates International Limited, 1982.

BARRETT, J.T.

1980 Basic immunology and its medical application (2^{TM*} edition). The C.V. Mosby Company, 1980.

CALADRO, J.J.

1971 *Rheumatoid arthritis.* Clinical Symposia, vol. 23(1) (1971), 32 pp. CIBA Pharmaceutical Company, 1971.

CALIN, A.

1983 Diagnosis and management of rheumatoid arthritis. The Addison-Wesley Clinical Practice Series. Addison-Wesley Publishing Company, Inc., 1983.

CLAYTON, L.T., *editor*

1985 Taber's cyclopedic medical dictionary (15** edition). F.A.Davis Company, 1985.

COHEN, A.S., *editor*

1975 Laboratory diagnostic procedures in the rheumatic diseases *(2nd* edition). Little, Brown and Company, 1975.

CÔTÉ, R.A., *et* a/., *editor*

1984 Systematized nomenclature of medicine. College of American Pathologists, 1984.

ISRAEL, R.A., *et* a/., *editors*

1980 International classification of diseases, *9th* Revision, Clinical Modification. *(2nd* edition) U.S. Department of Health and Human Services, 1980.

KOFFLER, D.

1979 *The immunology of rheumatoid diseases.* Clinical Symposia, Vol. 31, No. 4, 36 pp. CIBA Pharmaceutical Company, 1979.

MCCARTY, D.J., *et* a/., *editors*

1979 Arthritis and allied conditions. A textbook of rheumatology *(9th* edition). Lea and Febiger, 1979.

MEDSGER, T.A., JR.

1975 *hit really arthritis? A guide to diagnosis.* Medical Opinion, vol. 4 (1975), pp. 14-21.

Miller, R.A., Pople, H.E., Jr., and Myers, J.D.

1982 *Internist-I, an experimental computer-based diagnostic consultant.for general internal medicine.* The New England Journal of Medicine, vol. 307 (1982), pp. 468-476.

Moll, J.M.H.

1983 **Management of rheumatic disorders.** Raven Press, 1983.

National Library of Medicine.

1984 **MEDICAL SUBJECT HEADINGS, TREE STRUCTURES, 1985.** National Library of Medicine, 1984.

Peterdorf, R.G., *et al., editors*

1983 **Harrison's principles of internal medicine (10th edition).** McGraw-Hill Book Company, 1983.

Rakel, R.E. *editor*

1985 **Conn's current therapy: Latest Approved Methods of Treatment for the Practicing Physician.** W.B.Saunders Company, 1985.

Rodnan, G.P., *et al., editors*

1983 **Primer on the rheumatic diseases (8th edition).** Arthritis Foundation, 1983.

Sheon, R.P., Moskowitz, R.W., and Goldberg, V.M.

1982 **Soft tissue rheumatic pain: Recognition, management, prevention.** Lea and Febiger, 1982.

Weed, L.L.

1971 **Medical records, medical education, and patient care. The problem-oriented record as a basic tool.** The Press of Case Western Reserve, 1971.

Natural Language Processing

Carbonell, J.G.

1978 *Towards a Self-Extending Parser.* In: **Proceedings of the 17th Meeting of the Association for Computational Linguistics.** 1979, pp. 3–7.

CARBONELL, J.G., BOGGS, W.M., MAULDIN, M.L., and ANICK, P.G.

1983 *The XCALIBUR Project, A Natural Language Interface to Expert Systems and Data Bases.* **In: Applications in Artificial Intelligence, edited by S. Andriole. Petrocelli Books Inc., 1985.**

CARBONELL, J.G. and HAYES, P.J.

1983 *Recovery Strategies for Parsing Extragrammatical Language.* American **Journal of Computational Linguistics, vol. 9 (1983), pp. 123-146.**

1986 *Natural Language Processing: Techniques and Applications.* In: **The Encyclopedia of Artificial Intelligence, edited by S. Shapiro. Wiley & Sons, 1986.**

HAYES, P.J., and MOURADIAN, G.V.

1980 *Flexible Parsing.* **In: Proceedings of the 18th Meeting of the Association for Computational Linguistics. 1980, pp. 97-103.**

PERRAULT, C.R., ALLEN, J.F., and COHEN, P.R.

1978 *Speech Acts as a Basis for Understanding Dialog Coherence.* In: **Proceedings of the Second Conference on Theoretical Issues in Natural Language Processing. 1978.**

SCHANK, R.C.

1975 Conceptual Information Processing. North-Holland Publishing Company, 1975.

STEINACKER, I., and TROST, H.

1983 *Structural relations - a case against case.* In: **Proceedings of the eighth International Joint Conference on Artificial Intelligence. 1983, vol. 2, pp. 627-629.**

GAZDAR, G., and PULLUM, G.K.

1985 *Computationally Relevant Properties of Natural Languages and Their Grammars.* **New Generation Computing, vol. 3 (1985), pp. 273-306.**